# MULTIPLE DISEASES PREDICTION USING RESTRICTED BOLTZMANN MACHINES

# MINI PROJECT REPORT

Submitted by

**Avinash.k** **REGISTER NO: 21TN0002**

**Raguram.R** **REGISTER NO: 21TN0018**

*In fulfilment of the requirement for the award of the Degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**PONDICHERRY UNIVERSITY**


**MANAKULA VINAYAGAR INSTITUTE OF TECHNOLOGY**
**KALITHEERTHALKUPPAM, DEPARTMENT OF INFORMATION TECHNOLOGY**

**PUDUCHERRY - 605 107.**

**April 2024**

# MANAKULA VINAYAGAR INSTITUTE OF TECHNOLOGY
# PONDIYCHERRY UNIVERSITY

## DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

## BONAFIDE CERTIFICATE

This is to certify that the project work entitled **"MULTIPLE DISEASES PREDICTION USING RESTRICTED BOLTZMAN MACHINES"** is a bonafide work done by **Avinash.k [REGISTER NO: 21TN0002], Raguram.R [REGISTER NO:21TN0018]**,in partial fulfillment of the requirement for the award of B.Tech Degree in **DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** by Pondicherry University during the academic year 2023-24.

| | |
|---|---|
| **PROJECT GUIDE** | **HEAD OF THE DEPARTMENT** |
| **Mrs. K. ANUPRIYA.,** | **MR.R.RAJBHARATH.,** |
| **ASSISTANT PROFESSOR** | **ASSOCIATE PROFESSOR&HOD** |

*Submitted for the University Examination held on.....*

**Internal Examiner** **External Examiner**

# ACKNOWLEDGEMENT

We express our deep sense of gratitude to **Theiva Thiru. N. Kesavan,** Founder, **Shri. M. Dhanasekaran,** Chairman & Managing Director**, Shri. S. V. Sugumaran,** ViceChairman, and **Dr. K. Narayanasamy** Secretary of **Sri Manakula Vinayagar Educational Trust, Puducherry** for providing the necessary facilities to successfully complete our mini project and report works.

We express our sincere thanks to our beloved Principal **Dr. S. Malarkkan** for having provided the necessary facilities and encouragement for the successful completion of this project work.

We express our sincere thanks to **MR,R.RAJBHARATH**, **Head of the Department**, Department of Artificial Intelligence and Machine Learning**,** for his support in making necessary arrangements for the conduction of the Project and for guiding us to execute our project successfully.

We express our sincere thanks to **Mrs. K. ANUPRIYA,** Assistant Professor, Department of Artificial Intelligence and Machine Learning for his consistent reviews which motivated us in completing a project.

We thank all our department faculty members, non-teaching staff, and my friends for helping us to complete the document successfully on time.

We sincerely express our thanks to our mini project **Mrs. K. ANUPRIYA,** Assistant Professor, Department of Artificial Intelligence and Machine Learning for continuously motivating and helping to develop our mini project.

We would like to express our eternal gratitude to our parents for the sacrifices they made for educating and preparing us for our future and their everlasting love and support. We thank the Almighty for blessing us with such wonderful people and for being with us always

**SUSTAINABLE DEVELOPMENT GOALS (SDGs) MAPPING**

**Title** : MULTIPLE DISEASES PREDICTION USING RESTRICTED BOLTZMAN MACHINES

**SDG Goal : SDG Goal-3 (Good Health and well being)**


**SDG Goal -3:**

Regular physical exercise has been shown to lower the risk of many chronic diseases, improve mental health, and increase general well-being. It is therefore one of the most important components of good health and wellbeing. To completely enjoy these advantages, people must, however, make sure that they exercise properly and securely. This is where the project's emphasis on adaptive exercise diligence comes in handy. The research enables individualised coaching and feedback by utilising machine learning algorithms to detect and monitor positions throughout exercise routines. This ensures that individuals do workouts with accuracy and efficiency while minimising the danger of accidents. Moreover, the adaptable nature of the project suggests that it can accommodate people with different degrees of fitness, talents, and medical issues. By adjusting training plans in response to real-time posture detection With monitoring, the initiative encourages accessibility and diversity, enabling people from all backgrounds to partake in physical exercise according to their needs and abilities. Thus, proactive health management and preventative care are promoted, which are essential elements of SDG

# ABSTRACT

The Multiple Diseases Prediction using Restricted Boltzmann Machines develop a predictive model leveraging RBMs for early disease detection. By analyzing diverse patient data, including symptoms, medical history, lifestyle factors, and possibly genetic information, RBMs extract intricate patterns to predict the likelihood of developing various diseases. The project involves data collection, preprocessing, and feature representation, followed by RBM model training and disease prediction. The model's performance will be evaluated using metrics like accuracy and precision, with validation on independent datasets. disease recognition utilizing Restricted Boltzmann Machines (RBMs) as the core predictive model. The objective is to develop a system capable of early detection and prediction of various diseases based on patient health data. Through comprehensive analysis of symptoms, medical history, lifestyle factors, and potentially genetic information, RBMs extract intricate patterns to determine the likelihood of developing specific diseases. Key phases of the project include data collection, preprocessing, and feature representation, culminating in RBM model training and disease prediction. Evaluation metrics such as accuracy and precision will gauge model performance, with validation against independent datasets. Anticipated outcomes encompass a robust predictive framework pinpointing significant risk factors and early indicators, thus empowering proactive healthcare interventions. Integration into clinical decision support systems promises enhanced diagnostic accuracy and informed treatment strategies, ultimately fostering improved patient outcomes and healthcare efficacy.

## TABLE OF CONTENT

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

The realm of healthcare is witnessing a paradigm shift towards preventative measures and early disease detection. Traditionally, diagnosis often occurs after the onset of symptoms, potentially delaying critical interventions. This project delves into the application of Restricted Boltzmann Machines (RBMs) for the development of a robust predictive model capable of identifying the risk of various diseases at an earlier stage. By leveraging diverse patient data, this system aims to empower proactive healthcare strategies and improve patient outcomes.

The escalating prevalence of chronic diseases underscores the critical need for enhanced disease prediction capabilities. Early detection allows for timely interventions, improving treatment efficacy and potentially mitigating the severity of the illness. However, conventional methods often rely on the manifestation of clinical symptoms, which may occur later in the disease progression. This project proposes a novel approach that utilizes RBMs to analyze a comprehensive dataset encompassing various aspects of patient health.

The healthcare landscape is experiencing a paradigm shift, emphasizing preventative measures and early disease detection. Traditional diagnosis often relies on overt symptoms, potentially delaying critical interventions. Early detection empowers healthcare professionals, potentially mitigating disease severity and improving treatment efficacy.

Chronic diseases like heart disease, Parkinson's disease, and diabetes pose a significant global burden. These conditions have devastating consequences if left undiagnosed and untreated. Early detection plays a crucial role in managing these chronic illnesses.

However, conventional methods for disease diagnosis often have limitations. Many rely on the presence of overt clinical symptoms, which may not manifest until the disease has progressed significantly. Additionally, some traditional diagnostic procedures can be invasive, expensive, or time-consuming.

The need for innovative approaches that can overcome these limitations and facilitate earlier disease detection is paramount. Machine learning (ML) and deep learning (DL) techniques offer promising avenues for addressing this challenge.

Machine learning encompasses algorithms that learn from data without explicit programming, identifying patterns for prediction. In healthcare, ML algorithms are used for disease risk assessment, patient stratification, and treatment recommendation.

Deep learning, a subfield of ML, utilizes complex artificial neural networks. These networks can learn intricate relationships from large amounts of data, demonstrating capabilities in image recognition, natural language processing, and time series forecasting. In healthcare, DL techniques are increasingly being explored for disease diagnosis and prediction, offering promising opportunities for early detection.

Restricted Boltzmann Machines (RBMs) are a specific type of artificial neural network employed in DL applications. RBMs possess a specific architecture with a single layer of visible units (representing input data) and a single layer of hidden units. These units are not directly connected within the same layer, but connections exist between the visible and hidden units.

Their strength lies in the ability to learn complex, hidden patterns within data sets, making them suitable for tasks like dimensionality reduction and feature extraction – crucial steps for disease prediction models. In essence, RBMs can act as a pre-processing stage, extracting meaningful features from raw data that can then be used by other ML algorithms to build robust disease prediction models.

The application of RBMs for disease prediction offers several advantages. Firstly, RBMs are adept at handling high-dimensional data, making them suitable for analyzing complex patient information including symptoms, medical history, lifestyle factors, and potentially even genetic data. Secondly, RBMs offer flexibility in their architecture, allowing for customization based on the specific disease prediction task. Additionally, RBMs can be employed as building blocks for more complex neural networks, creating powerful predictive models for multiple diseases.

This project delves into the potential of applying RBMs for early disease detection of three prevalent conditions: heart disease, Parkinson's disease, and diabetes. By leveraging the capabilities of RBMs

to analyze diverse patient data, this project aims to develop a predictive model capable of identifying individuals at risk for developing these diseases at an earlier stage.

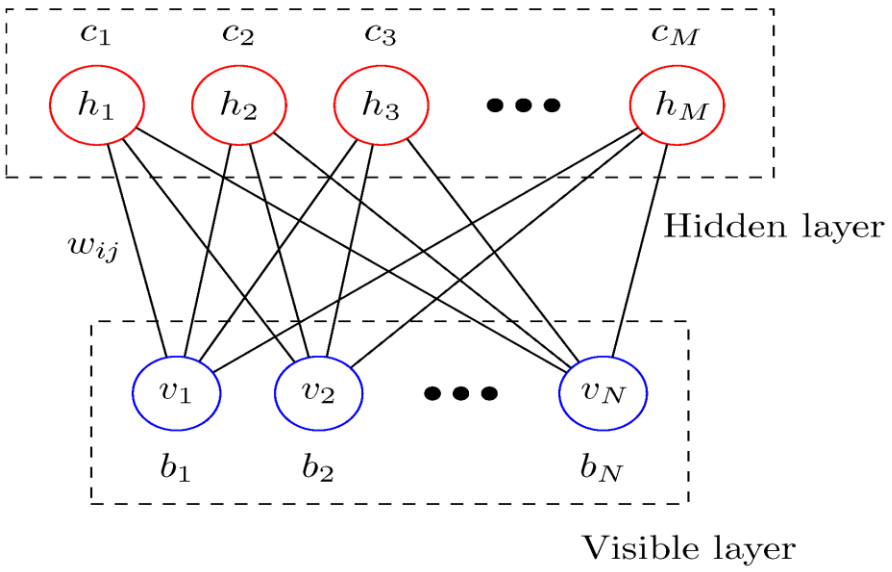



**fig. 1.1 Restricted Boltzmann Machines**

RBMs, a type of artificial neural network, excel at uncovering hidden patterns in complex data. This project capitalizes on this strength by employing RBMs to extract intricate relationships within patient information, including:

- **Symptoms:** The presence or absence of specific symptoms can provide valuable clues regarding potential underlying conditions.
- **Medical History:** Past medical records offer insights into a patient's susceptibility to certain diseases and the course of previous illnesses.
- **Lifestyle Factors:** Habits such as diet, exercise, and smoking can significantly influence disease risk.
- **Genetic Information (Optional):** When available, incorporating genetic data can further enhance the model's ability to identify individuals with predisposition to specific diseases.

Through a multifaceted analysis of this comprehensive data, the RBMs will learn to recognize subtle patterns that may foreshadow the development of various diseases. This project encompasses several key phases

**Table 1.1 Restricted Boltzmann Machine (RBM) vs Deep Boltzmann Machine (DBM)**

| Feature | Restricted Boltzmann Machine (RBM) | Deep Boltzmann Machine (DBM) |
|---|---|---|
| Visible Layer Connections | No connections between visible units | No connections within visible or hidden layers |
| Hidden Layer Connections | No connections between hidden units | Connections allowed between hidden units |
| Model Architecture | Two-layered (visible and hidden) | Multi-layered (visible, multiple hidden layers) |
| Learning | Focuses on learning the distribution of the visible data | Learns a hierarchical representation of the data |
| Training | Easier to train due to simpler architecture | More complex training due to deeper architecture |
| Applications | Feature extraction, dimensionality reduction, recommender systems | Feature learning, image recognition, natural language processing |

1. **Data Collection:** The initial stage involves the meticulous gathering of patient data from diverse sources, ensuring its accuracy, completeness, and ethical acquisition.
2. **Preprocessing:** The collected data may require cleaning and formatting to transform it into a suitable structure for RBM training. This may involve addressing missing values, handling inconsistencies, and potentially scaling the data for optimal performance.
3. **Feature Representation:** The raw data may require further processing to extract meaningful features that effectively capture the underlying relationships relevant to disease risk. Feature

engineering techniques may be employed to create informative representations for the RBM model.

4. **RBM Model Training:** The preprocessed data is then utilized to train the RBM. During this stage, the RBM learns to identify the intricate patterns within the data that hold predictive power for disease risk assessment.
5. **Disease Prediction:** Once trained, the RBM model is then used to analyze new patient data. Based on the learned patterns, the model predicts the likelihood of developing specific diseases for each individual patient.

Evaluating the model's performance is crucial to ensure its effectiveness. Metrics such as accuracy and precision will be employed to assess the model's ability to correctly identify individuals at risk for developing a particular disease. Furthermore, validation using independent datasets will enhance the model's generalizability and robustness.

The anticipated outcome of this project is a robust predictive framework that can pinpoint significant risk factors and early disease indicators. This information will empower healthcare professionals to implement proactive interventions, potentially preventing or mitigating the severity of diseases.

The integration of this RBM-based model into clinical decision support systems holds immense promise. By providing physicians with a data-driven assessment of disease risk, the system can facilitate informed treatment strategies and potentially enhance diagnostic accuracy. Ultimately, this project aims to contribute to improved patient outcomes and a more efficient healthcare system.

## 1.1 OVERVIEW

The healthcare landscape is undergoing a significant transformation, with a growing emphasis on preventative measures and early disease detection. Traditionally, diagnosis often occurs after the onset of symptoms, potentially delaying critical interventions. However, early detection empowers healthcare professionals to take proactive steps, potentially mitigating disease severity and improving treatment efficacy for chronic conditions like heart disease, Parkinson's disease, and diabetes.

These three prevalent conditions pose a significant global burden, impacting millions of individuals and carrying devastating consequences if left undiagnosed and untreated. Early detection plays a crucial role in managing these chronic illnesses. For instance, identifying heart disease early allows

for interventions that can prevent heart attacks and strokes, leading to improved cardiovascular health. Similarly, early diagnosis of Parkinson's disease facilitates the initiation of treatment strategies that manage symptoms and slow disease progression. Likewise, early detection of diabetes enables lifestyle modifications and medication management to prevent complications.

However, conventional methods for disease diagnosis often have limitations. Many methods rely on the presence of overt clinical symptoms, which may not manifest until the disease has progressed significantly. Additionally, some traditional diagnostic procedures can be invasive, expensive, or time-consuming. This highlights the need for innovative approaches that can overcome these limitations and facilitate earlier disease detection.

Machine learning (ML) and deep learning (DL) techniques offer promising avenues for addressing this challenge. Machine learning encompasses algorithms that learn from data without explicit programming, identifying patterns for prediction. In healthcare, ML algorithms are used for disease risk assessment, patient stratification, and treatment recommendation.

Deep learning, a subfield of ML, utilizes complex artificial neural networks. These networks can learn intricate relationships from large amounts of data, demonstrating capabilities in image recognition, natural language processing, and time series forecasting. In healthcare, DL techniques are increasingly being explored for disease diagnosis and prediction, offering promising opportunities for early detection.

This project delves into the potential of applying Restricted Boltzmann Machines (RBMs), a specific type of artificial neural network employed in DL applications. RBMs possess a unique architecture with a single layer of visible units representing input data (patient information) and a single layer of hidden units. These units are not directly connected within the same layer, but connections exist between the visible and hidden units.

The strength of RBMs lies in their ability to learn complex, hidden patterns within data sets. This makes them particularly well-suited for tasks like dimensionality reduction and feature extraction – crucial steps for disease prediction models. In essence, RBMs can act as a pre-processing stage,

and potentially even genetic data) that can then be used by other ML algorithms to build robust disease prediction models.

**Table 1.2 Diabetes, Heart Disease and Parkinson's Disease**

| Feature | Diabetes | Heart Disease | Parkinson's Disease |
|---|---|---|---|
| Type | Chronic metabolic disorder | Chronic condition affecting the heart and blood vessels | Neurodegenerative disorder affecting the nervous system |
| Main Cause | Impaired insulin production or function (Type 1) or insulin resistance (Type 2) | Buildup of plaque in arteries (atherosclerosis) | Loss of dopamine-producing neurons in the brain |
| Symptoms | Frequent urination, increased thirst, excessive hunger, unexplained weight loss, blurred vision | Chest pain, shortness of breath, pain radiating to arm/jaw/shoulder, fatigue, nausea, sweating | Tremor, rigidity, slowness of movement, difficulty with balance and coordination, speech problems |
| Complications | Heart disease, stroke, kidney disease, nerve damage, vision problems, foot ulcers | Heart attack, stroke, chest pain (angina), heart failure, arrhythmia | Dementia, depression, anxiety, sleep problems, swallowing difficulties |

The application of RBMs for disease prediction offers several advantages. Firstly, RBMs are adept at handling high-dimensional data, making them suitable for analyzing complex patient information. Secondly, RBMs offer flexibility in their architecture, allowing for customization based on the specific disease being predicted. Additionally, RBMs can be employed as building blocks for more complex neural networks, creating powerful predictive models for multiple diseases.

This project focuses on leveraging the capabilities of RBMs to develop a predictive model capable of identifying individuals at risk for developing heart disease, Parkinson's disease, and diabetes at an earlier stage. By analyzing diverse patient data, the model aims to empower healthcare professionals with the ability to take proactive steps towards disease prevention and improved patient outcomes.

### 1.1.1 AIMS AND OBJECTIVES

**Aim:**

The overall aim of this project is to develop a robust predictive model utilizing Restricted Boltzmann Machines (RBMs) for early detection of three prevalent diseases: heart disease, Parkinson's disease, and diabetes.

**Objectives:**

1. **Data Collection and Preprocessing:**
    - Acquire patient data from relevant sources, ensuring accuracy, completeness, and ethical considerations.
    - Preprocess the data to address missing values, inconsistencies, and potentially scale the data for optimal RBM performance.
2. **Feature Representation:**
    - Develop strategies to extract meaningful features from the raw patient data that effectively capture underlying relationships relevant to disease risk. This may involve feature engineering techniques suitable for the RBM model.
3. **RBM Model Training:**
    - Train the RBM model using the preprocessed data and optimized hyperparameters. During this stage, the RBM learns to identify the intricate patterns within the data that hold predictive power for disease risk assessment.
4. **Disease Prediction:**
    - Once trained, utilize the RBM model to analyze new patient data. Based on the learned patterns, predict the likelihood of developing specific diseases for each individual patient.
5. **Model Evaluation:**

Evaluate the model's performance using metrics like accuracy and precision to assess its ability to correctly identify individuals at risk for developing a particular disease. Validate the model's generalizability and robustness with independent datasets.

6. **Outcomes and Future Directions:**
    - Develop a robust predictive framework that pinpoints significant risk factors and early disease indicators.
    - Explore the potential integration of the RBM model into clinical decision support systems to enhance diagnostic accuracy and inform treatment strategies.
    - Identify future research directions, such as incorporating additional data sources or refining the RBM architecture for potentially improved model performance.

## 1.2WORKING PRINCIPLE

Restricted Boltzmann Machines (RBMs) offer a unique approach to analyzing complex patient data for early disease detection. These deep learning neural networks possess a distinct architecture with two key layers: a visible layer and a hidden layer.

The visible layer acts as the entry point for patient information, encompassing symptoms, medical history, lifestyle factors, and potentially even genetic data. Each element has a corresponding unit in this layer.

The hidden layer plays a crucial role in pattern recognition. Unlike the visible layer, units within the hidden layer are not directly connected to each other. However, connections exist between individual visible and hidden units. This restricted connectivity allows RBMs to focus on learning intricate relationships between different data points without influence from neighboring units within the same layer.

The core functionality of RBMs lies in their ability to learn the underlying probability distribution of the input data through a training algorithm called Contrastive Divergence (CD). This process involves presenting a real data point (patient information) to the visible layer, followed by activating hidden units based on the connections and associated weights.

The RBM then attempts to reconstruct the original data point based on the activated hidden units. This reconstruction represents the model's current understanding of the patient information.

Finally, the RBM utilizes the reconstructed data point to activate a new set of hidden units. The discrepancy between the original and reconstructed data is used to adjust the weights between the visible and hidden layers, iteratively refining the RBM's understanding of the data.

Through this probabilistic reconstruction and weight adjustment, the RBM effectively learns to extract meaningful features (hidden patterns) within the data that hold predictive power for disease risk assessment. These patterns might represent specific combinations of symptoms or medical history elements indicative of an increased risk for developing a particular disease.

Once trained, the RBM model can be used to analyze new patient data. Based on the activation patterns of the hidden layer units, the model predicts the likelihood of the patient developing a specific disease. This prediction leverages the learned relationships between the data points and disease risk captured during the training phase.

The advantages of RBMs for early disease prediction include their ability to handle high-dimensional data sets, their flexibility for customization based on the disease, and their potential as building blocks for complex multi-disease prediction models. By harnessing the power of RBMs to uncover hidden patterns within patient data, this project aims to develop a robust predictive model for early disease detection, empowering healthcare professionals to intervene at a critical juncture and ultimately improve patient outcomes.

**The Working Principle of RBMs for Early Disease Prediction**

This project delves into the potential of Restricted Boltzmann Machines (RBMs) for early disease detection. RBMs, a specific type of artificial neural network employed in Deep Learning (DL), offer a unique approach to analyzing complex patient data and identifying individuals at risk for developing diseases like heart disease, Parkinson's disease, and diabetes. Here, we explore the working principle of RBMs and how they can be harnessed for early disease prediction.

**Table 1.3 Machine Learning Algorithm VS Restricted Boltzmann Machine (RBM)**

| Feature | Machine Learning Algorithm | Restricted Boltzmann Machine (RBM) |
|---|---|---|
| Type | General category of algorithms | Specific type of unsupervised learning algorithm |
| Learning Style | Can be supervised, unsupervised, or reinforcement | Unsupervised |
| Goal | Wide range of goals, including classification, regression, dimensionality reduction. | Feature extraction, dimensionality reduction |
| Layers | Can have various layer structures, including single layer, multi-layer, or recurrent | Two layers: visible and hidden |
| Connections | Connections can exist within and between layers (depending on the algorithm) | No connections within visible or hidden layers (restricted) |
| Training | Trained using various methods depending on the algorithm's goal | Trained to minimize a specific energy function |
| Applications | Broad range of applications in various domains | Feature learning for deep learning architectures, anomaly detection, recommender systems |

**Understanding the Architecture:**

RBMs possess a distinct architecture characterized by two layers: a visible layer and a hidden layer. The visible layer serves as the entry point for patient data. This data can encompass a variety of elements relevant to disease risk assessment, such as symptoms, medical history, lifestyle factors, and potentially even genetic information. Each element in the data is represented by a single unit in the visible layer.

The hidden layer, on the other hand, plays a crucial role in pattern recognition. Unlike the visible layer, the hidden units are not directly connected to each other. However, connections exist between

individual units in the visible layer and individual units in the hidden layer. This restricted connectivity is a defining characteristic of RBMs, allowing them to focus on learning intricate relationships between different data points without the influence of neighbouring units within the same layer.

**Learning Through Probabilistic Reconstruction:**

The core functionality of RBMs lies in their ability to learn the underlying probability distribution of the input data. This learning process is achieved through a training algorithm called Contrastive Divergence (CD). Here's how it works:

1. **Positive Phase:** The training process begins with a "positive phase" where a real data point representing a patient's information is presented to the visible layer units. Based on the connections and associated weights between the visible and hidden layers, the RBM activates certain hidden units with varying probabilities. This activation pattern reflects the RBM's initial understanding of the relationships within the data.
2. **Reconstruction:** In the subsequent step, the RBM attempts to reconstruct the original data point based on the activated hidden units. It does this by calculating the probability of each visible unit being activated given the state of the hidden units. This reconstructed data point represents the RBM's current estimate of the original patient information.
3. **Negative Phase:** The final step involves a "negative phase" where the RBM utilizes the reconstructed data point to activate a new set of hidden units. This activation pattern reflects the RBM's understanding of the reconstructed data, which may differ from the original data.
4. **Weight Adjustment:** The key to learning lies in the discrepancy between the original data and the reconstructed data. The RBM compares the activation states of the visible units in both phases and adjusts the weights between the visible and hidden layers to minimize this discrepancy. By iterating through these positive and negative phases with multiple data points, the RBM progressively refines its understanding of the underlying relationships within the patient data.

**Feature Extraction: Unveiling the Hidden Landscape** Through this process of probabilistic reconstruction and weight adjustment, the RBM effectively learns to extract meaningful features from the complex patient data. These features represent the hidden patterns within the data that hold

predictive power for disease risk assessment. For instance, the RBM might identify specific combinations of symptoms or medical history elements that are indicative of an increased risk for developing heart disease.

**Building the Predictive Model:**

Once trained, the RBM model can be employed to analyze new patient data. The same process of activation and reconstruction is performed on the new data point. Based on the activation patterns of the hidden layer units, the model can predict the likelihood of the patient developing a specific disease. This prediction is based on the learned relationships between the data points and the disease risk, which were captured during the training phase.

### 1.2.1Identification

Early disease detection is crucial for improved health outcomes. Traditional methods often have limitations. Machine learning offers promising avenues, and this project explores Restricted Boltzmann Machines (RBMs). RBMs can analyze complex patient data to identify hidden patterns linked to disease risk. By leveraging RBMs, this project aims to develop a model for early detection of heart disease, Parkinson's disease, and diabetes.

### 1.2.2Verification

The proposed model utilizing RBMs requires verification. This involves evaluating its performance on unseen patient data. Metrics like accuracy and precision will assess its ability to correctly identify individuals at risk for specific diseases. Additionally, comparing results with existing methods will determine the effectiveness of the RBM approach for early disease detection.

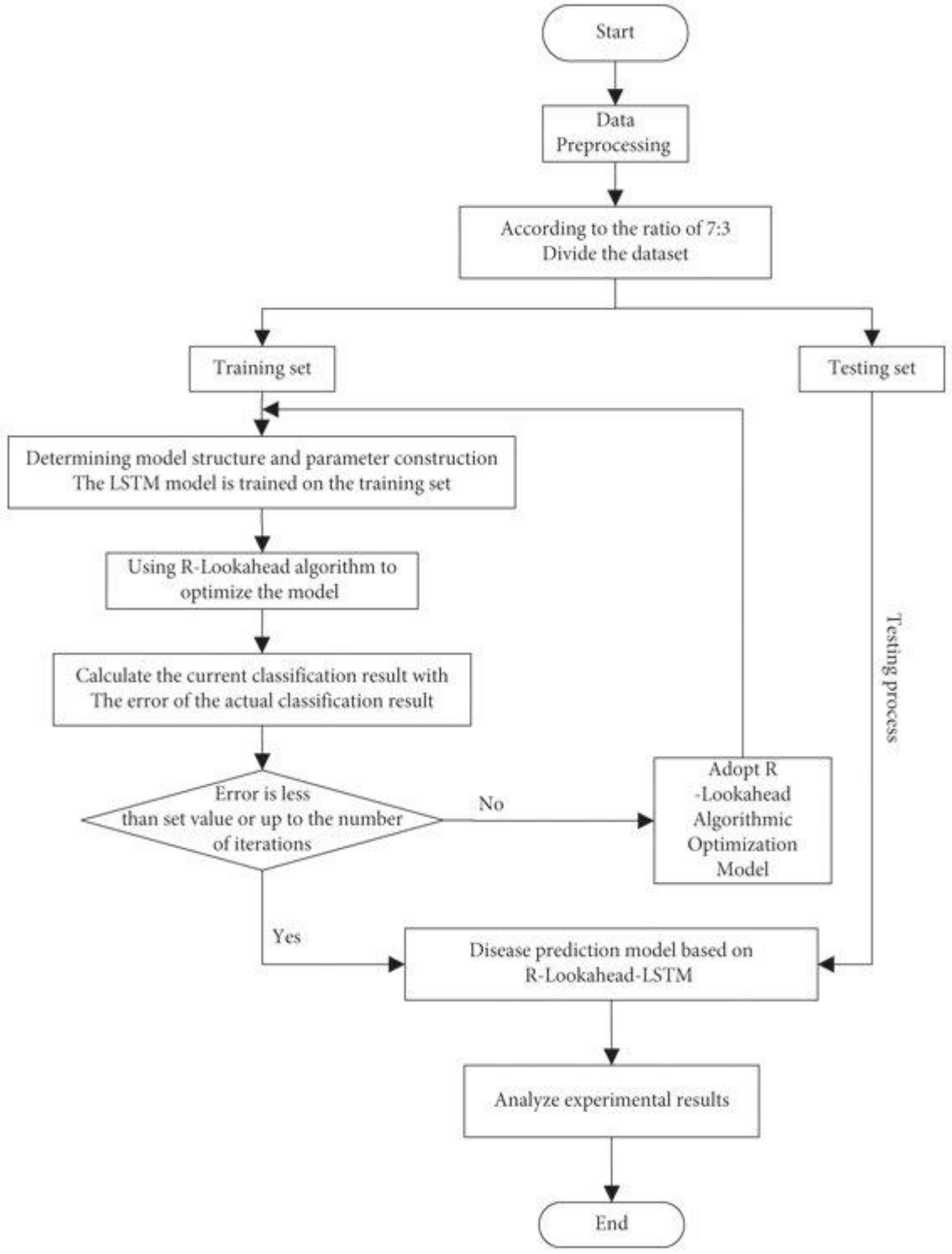



**Fig. 1.2 Flow diagram for training dieseas prediction model based on LSTM**

## Description

- **Data Collection (Start):** The process begins by gathering data relevant to the disease you want to predict (e.g., patient demographics, medical history, lab test results).
- **Data Preprocessing (Block labeled "Data Preprocessing"):** The collected data is cleaned and prepared for use in the RBM model. This might involve handling missing values, normalization, or feature engineering.
- **Divide Data into Sets (Block labeled "Data Preprocessing" with arrow pointing down):** The data is divided into two sets: a training set used to train the RBM model and a testing set used to evaluate its performance on unseen data. The ratio of this split is likely indicated somewhere on the flowchart (not mentioned in your description).
- **RBM Training** (Blocks labeled "Training set" and "LSTM model"): The RBM model is trained on the training set. Here's a general idea of the training process:
- An LSTM model (a type of recurrent neural network) is used within the RBM for training. This LSTM likely helps the RBM capture sequential information in the data, which might be useful for disease prediction tasks.
- **Look-ahead Optimization (Block labeled "Using R-Lookahead algorithm to optimize the model"):** An R-Lookahead algorithm is used for optimization during training. This algorithm considers future states during weight updates, potentially improving the convergence and performance of the RBM model.
- **Error Calculation and Stopping Criteria (Block labeled "Calculate the current classification result with The error of the actual classification result" with arrows):** During training, a block calculates the error between the model's predictions and the actual data. The flowchart likely includes a decision point (not mentioned in your description) based on this error. If the error is below a certain threshold or a maximum number of training iterations is reached, the training process stops. Otherwise, it loops back to step 4.
- **Testing** (Block labeled "Testing process"): Once the RBM model is trained, it's evaluated on the testing set. This assessment helps gauge how well the model generalizes to unseen data.
- **Disease Prediction (Block labeled "Disease prediction model based on R-Lookahead-LSTM"):** If the model performs well on the testing set, it can be used for disease prediction.

# CHAPTER 2

## EXISTING SYSTEM

Traditional methods for diagnosing diseases like heart disease, Parkinson's disease, and diabetes often rely on the presence of overt symptoms. However, these symptoms may not manifest until the disease has progressed significantly. Additionally, some diagnostic procedures like coronary angiography for heart disease can be invasive and expensive.

Existing approaches to diagnosing diseases like heart disease, Parkinson's, and diabetes often rely on readily apparent symptoms, which may appear late in the disease course. Additionally, some procedures like coronary angiography can be invasive and expensive.

Advancements in imaging and other technologies exist, but they're typically used after symptom presentation. Machine learning offers a promising alternative for earlier detection. Current ML systems for disease prediction involve algorithms trained on historical data to identify patterns and risk factors.

These systems might utilize Support Vector Machines (SVMs) or Random Forests for disease classification based on patient data. Others explore deep learning techniques like Convolutional Neural Networks (CNNs) to analyze medical images for disease-related abnormalities.

While advancements exist in imaging and other technologies, they're often employed after symptom presentation. Machine learning (ML) offers a promising alternative for earlier detection. Existing ML approaches for disease prediction typically involve algorithms trained on historical data to identify patterns and risk factors.Some existing systems utilize Support Vector Machines (SVMs) or Random Forests for disease classification based on patient data. Others explore deep learning techniques like Convolutional Neural Networks (CNNs) for analyzing medical images to detect abnormalities indicative of disease.

However, these existing systems may have limitations in handling high-dimensional data encompassing various aspects like symptoms, medical history, and potentially even genetic information. Additionally, some ML algorithms require large datasets for optimal performance, which can be a challenge in certain healthcare settings.

# CHAPTER 3

# PROPOSED SYSTEM

Current methods for diagnosing heart disease, Parkinson's disease, and diabetes often rely on symptoms that appear late in the disease progression. Additionally, traditional procedures can be invasive and expensive.

This project proposes a novel system utilizing Restricted Boltzmann Machines (RBMs), a powerful deep learning technique. Unlike traditional methods, this system aims to identify individuals at risk for these diseases at an earlier stage.

The system will collect patient data encompassing symptoms, medical history, lifestyle factors, and potentially even genetic information. This data will be preprocessed and transformed to extract meaningful features suitable for the RBM model.

The core of the system lies in the RBM, which will be trained on the processed data. During training, the RBM learns to identify intricate patterns within the data that hold predictive power for disease risk assessment.

Once trained, the model can analyze new patient data. Based on the learned patterns, the model will predict the likelihood of developing specific diseases for each individual. This allows for earlier detection compared to traditional methods.

The proposed RBM system offers several advantages. Firstly, RBMs excel at handling complex data sets, making them suitable for analyzing diverse patient information. Secondly, the RBM architecture can be customized based on the specific disease being predicted. Finally, by identifying hidden patterns linked to disease risk, the system aims for earlier detection, potentially leading to improved patient outcomes.

This project focuses on developing a robust predictive model with RBMs for early disease detection, empowering healthcare professionals with the ability to take proactive steps towards disease prevention.

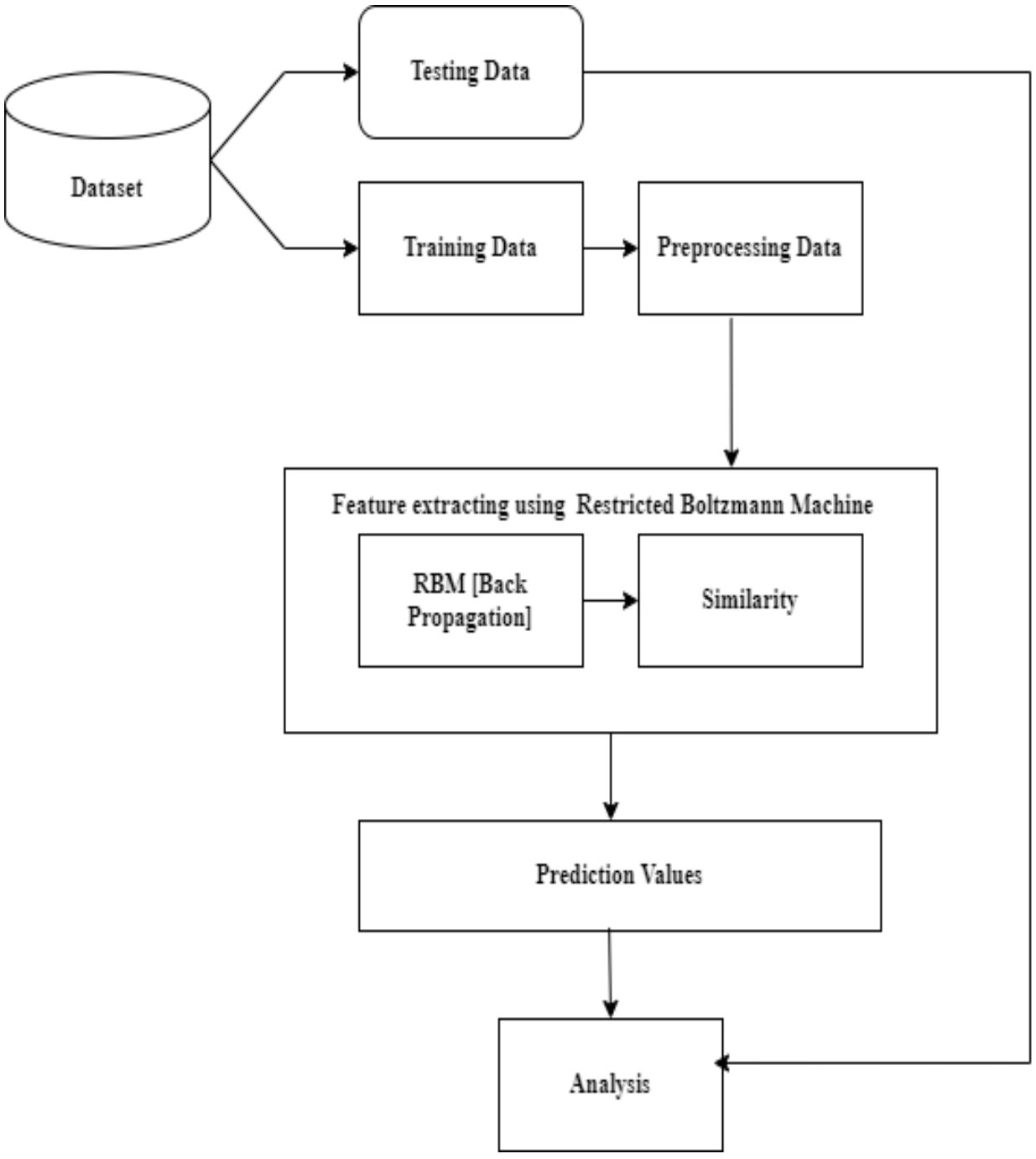



**Fig. 3.1 Architecture diagram for feature extracting using Restricted Boltzmann Machine**

## Description

**Disease Prediction Using Restricted Boltzmann Machine (RBM)**

1. **Data Collection and Preprocessing** (Start box): The process begins by gathering data relevant to the disease you want to predict (e.g., patient demographics, medical history, lab test results). This data is then cleaned and prepared for use in the RBM model.
2. **Define RBM Architecture** (Block labeled "Define RBM"): This step specifies the structure of the RBM, including the number of visible units (representing input features) and hidden units (representing features learned by the RBM).
3. **RBM Training** (Blocks labeled "Positive Phase" and "Reconstruction Phase"): The RBM is trained through an iterative process involving two main phases:
    - **Positive Phase:** The activation probabilities of hidden units are calculated based on the input data (visible layer activations).
    - **Reconstruction Phase:** The visible layer is reconstructed based on the activated hidden units.
4. **Weight Update** (Block labeled "Contrastive Divergence" with an arrow pointing to "Update Weights & Biases"): The weights and biases within the RBM are updated after each iteration. This update minimizes the difference between the original data and its reconstruction, allowing the RBM to learn the underlying patterns in the data.
5. **Optional Fine-tuning** (Decision diamond labeled "Fine-tuning"): This step allows for optionally refining the model for disease classification. It might involve using the hidden layer activations (extracted features) from the RBM as input to a supervised learning algorithm like Logistic Regression, trained on labeled data (patients with or without the disease).
6. **Disease Prediction** (Block labeled "Disease Prediction"): Once the model is trained (RBM or fine-tuned supervised model), new patient data can be fed into the model to predict the presence of the disease.
7. New patient data may be incorporated into the trained model (either an RBM or an improved supervised model) to predict the existence of the illness.
8. After every iteration, the RBM's weights and biases are changed. By reducing the discrepancy between the original and reconstructed data, this update enables the RBM to discover the underlying patterns in the data.

# CHAPTER 4

# REQUIREMENTS

## Detailed Requirements for RBM-based Early Disease Detection System

This section outlines the hardware, software, libraries, and data set requirements for developing a robust predictive model utilizing Restricted Boltzmann Machines (RBMs) for early disease detection of heart disease, Parkinson's disease, and diabetes.

## 4.1 HARDWARE REQUIREMENT

- **Processor**
  - **Minimum:** A multi-core processor (minimum 4 cores) with a clock speed of at least 3.0 GHz is recommended for efficient data processing. Consider processors with hyperthreading capabilities to further improve performance.
  - **Optimal:** For handling larger and more complex datasets, a high-performance multi-core processor (minimum 8 cores) with clock speeds exceeding 3.5 GHz is ideal. Look for CPUs with support for vector instruction sets like AVX or AVX2, which can significantly accelerate specific deep learning operations.
- **Memory (RAM)**
  - **Minimum:** A minimum of 16 GB of RAM is essential to handle the demands of data loading, preprocessing, and model training, especially if dealing with high-dimensional patient data sets.
  - **Optimal:** Consider allocating 32 GB or more of RAM for smoother operation and the ability to work with larger datasets or more complex RBM architectures.
- **Storage**
  - **Type:** A solid-state drive (SSD) with sufficient storage capacity is crucial for fast data loading and model training. SSDs offer significantly faster read/write speeds compared to traditional hard disk drives (HDDs).
  - **Capacity:** The storage capacity will depend on the size of your chosen dataset and potential for future expansion. Aim for at least 500 GB of storage, with 1 TB or more

recommended for larger datasets or the addition of pre-trained models or intermediate results.

- **Graphics Processing Unit (GPU) (Optional)**
    - While not strictly necessary, a dedicated GPU can significantly accelerate the training process of RBM models, especially with larger datasets. GPUs are particularly adept at handling the parallel processing tasks involved in deep learning.
    - **Specifications:** Look for GPUs with high memory capacity (VRAM) suitable for deep learning tasks. A minimum of 8 GB of VRAM is recommended, with 16 GB or more preferred for more complex models or larger datasets. Consider NVIDIA's GeForce RTX series or AMD's Radeon RX series for deep learning compatibility.

## 4.2 SOFTWARE REQUIREMENT

- **Operating System**
    - A stable operating system like Linux (e.g., Ubuntu LTS version) is commonly preferred for deep learning projects due to its open-source nature, flexibility, and compatibility with various libraries. Linux distributions like Ubuntu offer pre-configured environments specifically designed for deep learning.
    - Windows 10 or macOS can also be used, but compatibility with specific libraries might need verification.
- **Python**
    - Python is the primary programming language for this project. Ensure you have a recent version of Python installed (e.g., Python 3.7 or later) with appropriate package management tools like pip for installing required libraries.
- **Deep Learning Framework**
    - Several deep learning frameworks support RBM implementation. Popular options include.
        - **TensorFlow:** A powerful and versatile framework with extensive documentation and community support. TensorFlow offers high-level abstractions and low-level control for building and training complex models.

- **PyTorch:** Another popular choice known for its ease of use, dynamic computational graph, and research-oriented features. PyTorch can be particularly efficient for prototyping and rapid development of deep learning models.

## 4.3 LIBRARIES

Depending on the chosen deep learning framework and specific functionalities you plan to implement, additional libraries might be required.

- **NumPy:** Provides efficient numerical computation capabilities essential for data manipulation and mathematical operations within deep learning models.
- **Scikit-learn:** Offers functionalities for data preprocessing tasks like normalization, feature scaling, and potential feature engineering relevant to RBM model input.
- **Matplotlib/Seaborn:** Used for data visualization, allowing you to explore your data set, analyze training progress, and present model results effectively.
- **Keras (Optional):** While often integrated with TensorFlow, Keras can be used as a high-level API for building and training deep learning models in a more user-friendly manner.

### Data Manipulation Libraries

- NumPy: For numerical computing and array operations.
- pandas: For data manipulation and analysis.
- SciPy: For scientific computing and statistical analysis.

### Machine Learning Libraries

- scikit-learn: For machine learning algorithms and model evaluation.
- TensorFlow or PyTorch: For deep learning models if needed for advanced analysis.

## 4.4 DATA SET

**Data Set Considerations for RBM-based Early Disease Detection System**

Obtaining a high-quality and comprehensive data set is crucial for developing a robust RBM model for early disease detection. Here's a breakdown of the data elements and considerations:

- **Data Sources**

There are several potential sources for patient data relevant to early disease detection:

**Electronic Health Records (EHR): EHR systems maintained by hospitals and clinics can be a rich source of patient information, including demographics, medical history, diagnoses, medications, lab results, and potentially even lifestyle factors documented during consultations. However, access to EHR data often requires collaboration with healthcare institutions and adherence to strict privacy regulations.**
**Public Health Databases: Government agencies or public health organizations may maintain anonymized health data sets that can be valuable for research purposes. These data sets may encompass demographics, disease prevalence statistics, and risk factors.**
**Clinical Trials and Research Studies: Data collected during clinical trials or research studies investigating specific diseases can offer valuable insights, particularly if the studies focus on early detection methods. Accessing such data might require contacting the research institutions or study leads.**
**Wearable Device Data: With growing adoption of wearable health trackers and smartwatches, anonymized data from these devices can potentially provide insights into lifestyle factors like physical activity, sleep patterns, and heart rate variability, which may be relevant to disease risk assessment. However, privacy concerns and potential limitations in data quality need to be considered.**

- **Data Description**

The data set should ideally include the following information for each patient:

**Unique Identifier: An anonymized identifier to link different data points for a particular patient while maintaining patient privacy.**
**Demographics: Age, gender, ethnicity, and potentially even socioeconomic factors (if ethically permissible and relevant to the diseases being studied).**
**Medical History: Past diagnoses of relevant diseases, medications taken, and any major medical procedures undergone.**
**Symptoms: Specific symptoms reported by the patient that may be indicative of each disease being targeted (heart disease, Parkinson's disease, diabetes). Standardizing symptom reporting or using established symptom scoring systems can improve data consistency.**
**Lifestyle Factors: Smoking habits, alcohol consumption, dietary habits, physical activity levels, and body mass index (BMI).**
**(Optional) Genetic Information: If available and ethically permissible, anonymized genetic data can potentially enhance model performance by identifying genetic markers associated with**

**disease risk. However, obtaining genetic data often requires stricter ethical considerations and informed patient consent.**

- **Data Quality Considerations**
    - **Completeness:** Ensure the data set has minimal missing values. Missing entries can significantly impact the training process and model performance. Techniques like data imputation or data exclusion might be necessary to address missing values.
    - **Consistency:** Standardize data formats and units of measurement across different data points. Inconsistent data can lead to errors during analysis and model training.
    - **Accuracy:** Verify the accuracy of the data through cross-checking with other sources or employing data cleaning techniques to identify and correct potential errors.
    - **Privacy:** Ensure the data set adheres to ethical guidelines and patient privacy regulations. Anonymize patient data and obtain proper authorization for data usage whenever necessary.
- **Data Size Considerations**
    - While a larger data set can potentially lead to a more robust model, consider computational limitations and available resources. Training RBMs on massive datasets can be resource-intensive, requiring powerful hardware and potentially longer training times.
    - Start with a well-curated medium-sized data set and gradually increase the size as your computational resources and model development progress.

A high-quality and curated data set containing relevant patient information is essential for training and validating the RBM model. The data set should ideally encompass the following:

- **Patient Demographics:** Age, gender, ethnicity, etc.
- **Medical History:** Past diagnoses, medications, relevant medical procedures, etc.
- **Symptoms:** Reported symptoms specific to each disease being targeted (heart disease, Parkinson's disease, diabetes).

# CHAPTER 5

# DESIGN COMPONENTS

1. **Data Acquisition and Preprocessing:** Data relevant to heart disease, Parkinson's disease, and diabetes will be collected from reliable sources. This data will then undergo cleaning and processing to address missing values, inconsistencies, and potentially be transformed to extract meaningful features suitable for the RBM model.
2. **RBM Model Design and Training:** The RBM architecture will be defined, specifying the number of units for representing patient data and capturing hidden relationships. Hyperparameters crucial for training will be optimized, and a training algorithm like Contrastive Divergence (CD) will be implemented to refine the model's understanding of the data.
3. **Disease Prediction and Evaluation:** Once trained, the model will analyze new patient data, activate hidden units based on the information, and predict the likelihood of specific diseases using learned patterns. Metrics like accuracy and precision will assess the model's ability to identify individuals at risk. Validation with a separate data set will ensure generalizability.
4. **Optional Components:** An API (Application Programming Interface) could be developed to integrate the model with existing systems, allowing healthcare professionals to access its prediction functionality. Additionally, a user interface could be designed to facilitate user interaction and data input for risk assessment.

The system will be designed with modularity, documentation, and scalability in mind for future maintenance, expansion, and adaptation to handle larger datasets or incorporate new functionalities.

This design approach utilizes RBMs to analyze complex patient data, identify hidden patterns linked to disease risk, and ultimately predict the likelihood of specific diseases at an earlier stage, empowering healthcare professionals to take proactive measures.

**Data Ingestion and Storage**

- Healthcare datasets from various sources are ingested into the system.

- Data preprocessing and cleaning are performed to handle missing values and inconsistencies.
- Cleaned data is stored in a relational or NoSQL database.

**Collaboration and Development Environment**

- Collaboration tools facilitate communication and task management among team members.
- Version control systems ensure the integrity and manageability of the codebase.
- Development environments provide tools for coding, testing, and debugging.

**Data Analysis and Visualization**

- Data analysis software processes healthcare datasets using statistical and machine learning techniques.
- Visualization tools create interactive dashboards and plots to present analysis results.
- Insights from the data are communicated to stakeholders for informed decision-making.

**Testing and Deployment**

- Testing frameworks validate the functionality and accuracy of the fuzzy DEA model.
- Deployment tools automate the deployment process into production or testing environments.
- Continuous integration and continuous deployment (CI/CD) pipelines ensure smooth deployment and updates workflow

**Data Preparation**

- Raw healthcare data is collected from diverse sources.
- Preprocessing steps include cleaning, normalization, and feature engineering.

**Model Development**

- Data scientists and fuzzy logic specialists collaborate to design and implement the fuzzy DEA model.
- Libraries like scikit-fuzzy and NumPy are utilized for fuzzy logic operations and numerical computations.

**Validation and Testing**

- The developed model is rigorously tested using real and simulated data.
- Performance metrics are evaluated to ensure the accuracy and reliability of the model.

**Pilot Testing**

- Pilot testing is conducted in a specific healthcare area to validate the model's effectiveness in real-world scenarios.
- Feedback from healthcare practitioners is incorporated to refine the model.

**Deployment and Monitoring**

- The validated model is deployed into production or testing environments.
- Monitoring tools track the model's performance and detect anomalies or drifts.
- Regular updates and maintenance ensure the model's relevance and accuracy over time.

Technologies and Tools:

- Programming Languages: Python, R
- Libraries: scikit-fuzzy, NumPy, pandas, Matplotlib, Seaborn, Plotly
- Database: PostgreSQL, MongoDB
- Collaboration Tools: Slack, Git
- Development Environments: Jupyter Notebook, PyCharm, RStudio
- Testing Frameworks: pytest, unit test
- Deployment Tools: Docker, Kubernetes, Jenkins

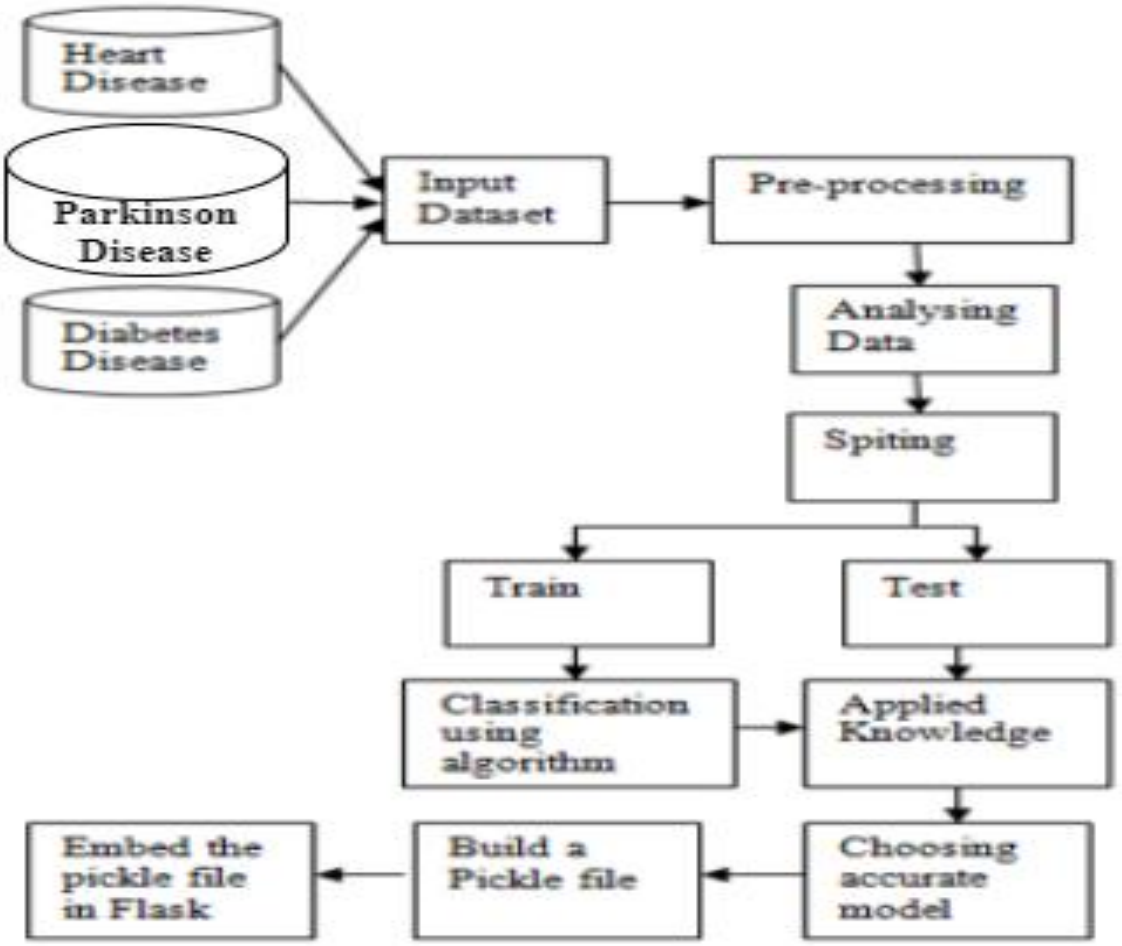



**fig. 5.1 State Diagram for disease prediction**

## Description

Based on the image you sent, it appears to be a flowchart depicting a process for disease prediction using a Restricted Boltzmann Machine (RBM) with a Long Short-Term Memory (LSTM) network and R-Lookahead optimization.

1. **Data Collection (Start):** The process starts by gathering relevant disease data (e.g., patient demographics, medical history, lab test results).
2. **Data Preprocessing (Block labeled "Data Preprocessing"):** The data is cleaned and prepared for the RBM model.
3. **Data Splitting** (Block labeled "According to the ratio of 7:3 / Divide the dataset"): The data is divided into training and testing sets (70% for training, 30% for testing).

4. **RBM Training** (Blocks labeled "Training set" and "LSTM model"): The RBM model with LSTM is trained on the training set. The LSTM likely helps capture sequential information in the data.
5. **Look-ahead Optimization** (Block labeled "Using R-Lookahead algorithm to optimize the model"): The R-Lookahead algorithm refines the model during training to improve convergence.
6. **Error Check & Stopping Criteria** (Block labeled "The error of the actual classification result"): The model's error is monitored. Training stops when the error is low enough or a maximum number of iterations is reached.
7. **Testing** (Block labeled "Testing process"): The trained model is evaluated on the testing set to assess generalizability.
8. **Disease Prediction** (Block labeled "Disease prediction model based on R-Lookahead-LSTM"): If the model performs well, it can be used for disease prediction on new data.

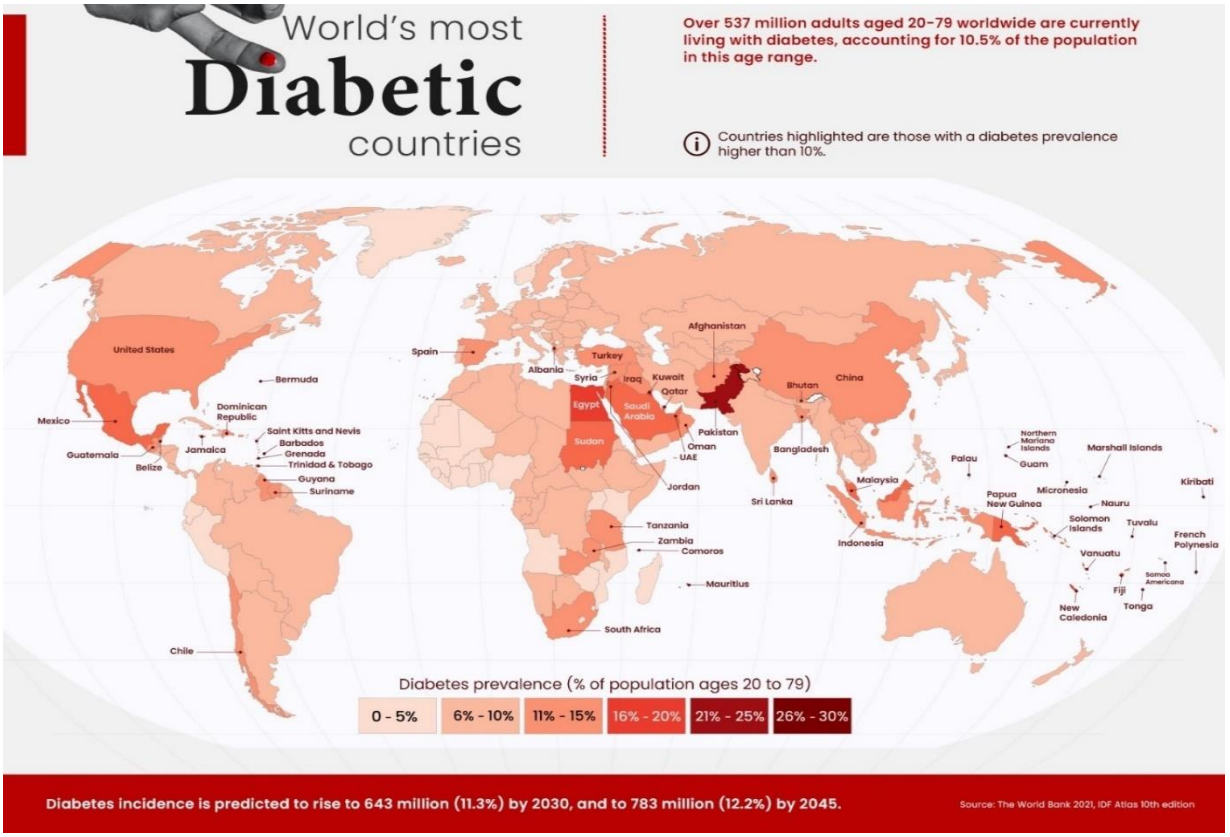



**fig. 5.2 The world Diabetics Survey**

## Description

highlights countries with a high prevalence of diabetes among adults aged 20 to 79.

- The title states that over 537 million adults aged 20-79 worldwide are currently living with diabetes, accounting for 10.5% of the population in this age range.
- The colors on the world map correspond to the prevalence of diabetes. Darker colors represent countries with a higher percentage of their population having diabetes.
- A legend at the bottom of the diagram explains the color coding which ranges from 0-5% (lightest color) to 26%-30% (darkest color).
- Several countries are highlighted, including the United States, Mexico, China, Saudi Arabia, and Egypt. Text boxes around the edges of the map list some of these highlighted countries.

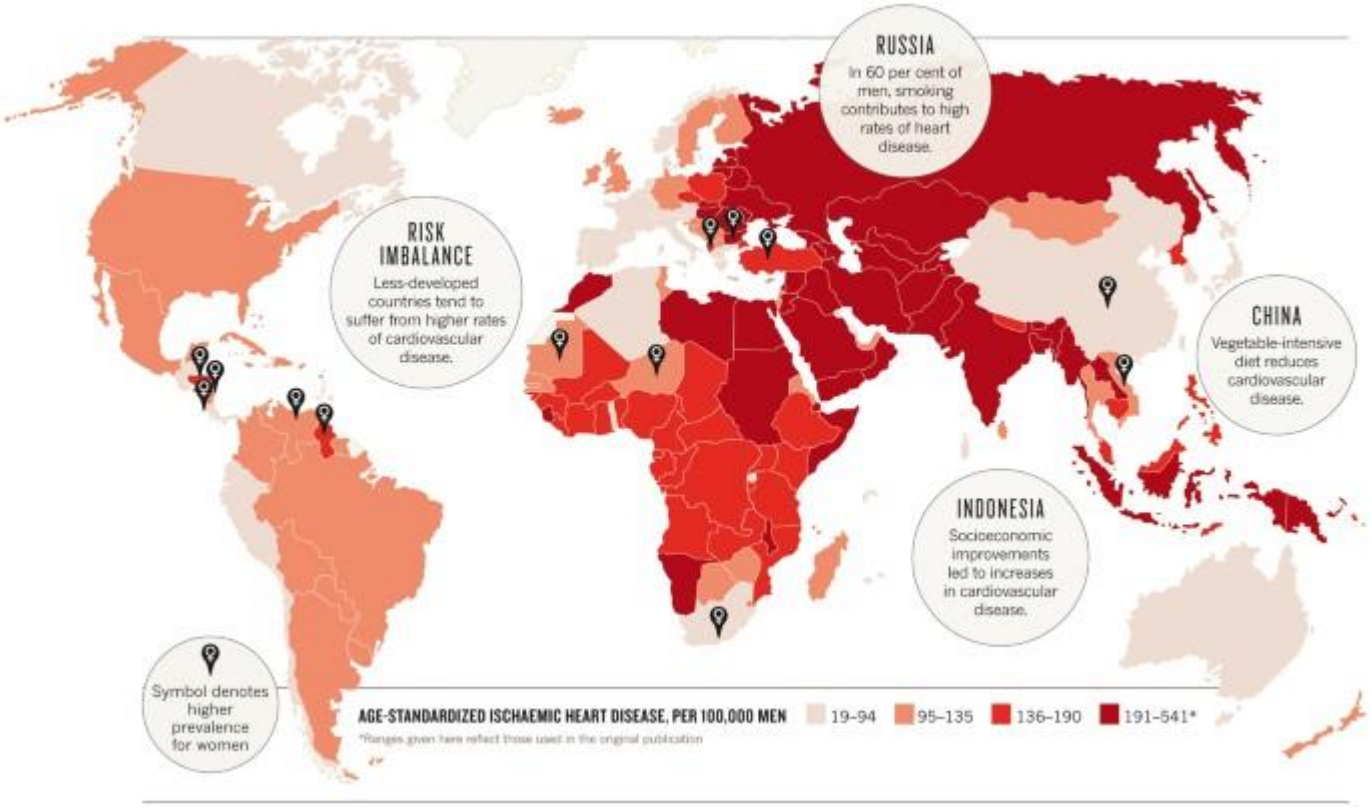



**fig. 5.3 The world heart disease survey**

## Description

**Gather & clean data** on patients.

**Define RBM structure** with visible and hidden units.

**Train the RBM** through positive and reconstruction phases.

**Update weights** to improve feature learning from data.

**Fine-tune** with a supervised model for disease classes.

**Predict disease** for new patients using the trained model.

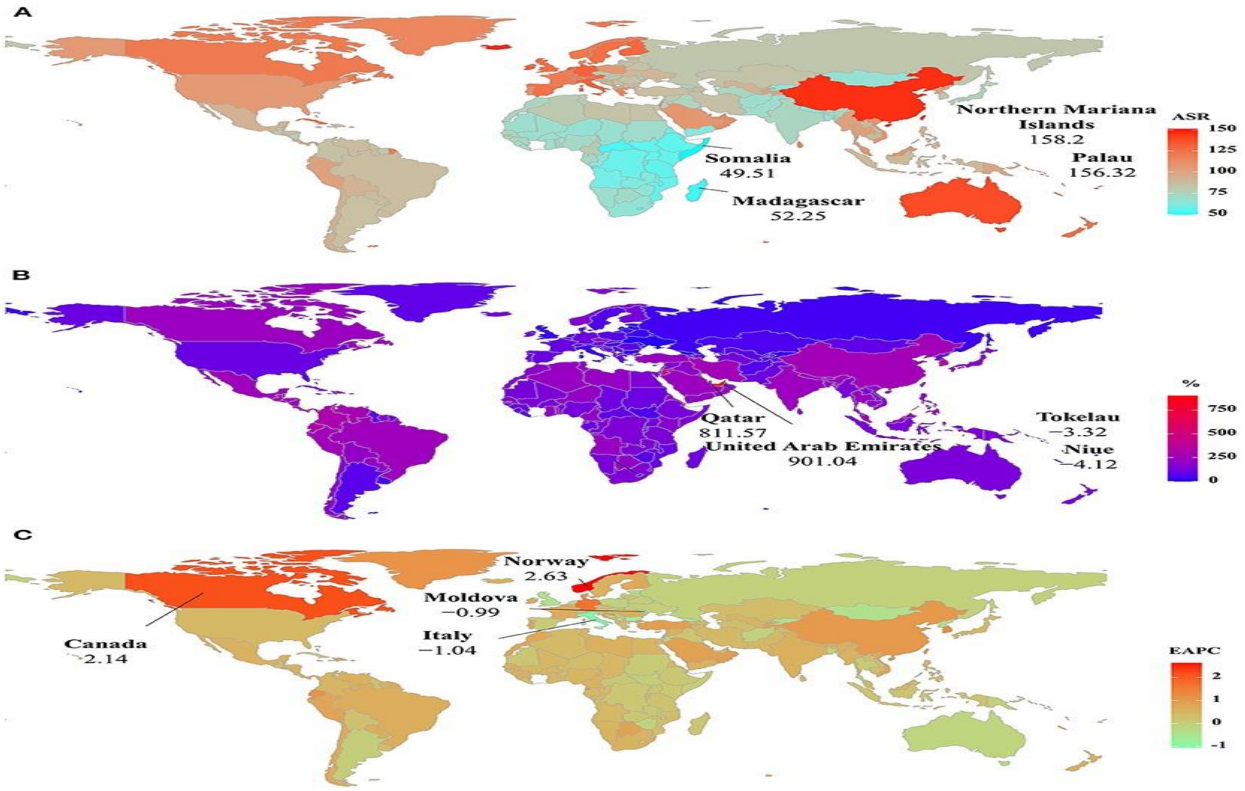



**fig. 5.4 The world parkinson disease survey**

## Description

**Collect & clean patient data**.

**Define RBM architecture** (visible & hidden units).

**Train RBM:** positive & reconstruction phases.

**Update weights** to improve feature learning.

**Fine-tune** with supervised model for disease classes.

**Predict disease** for new patients using the model.

**(Implicit):** Trained model captures relevant disease patterns

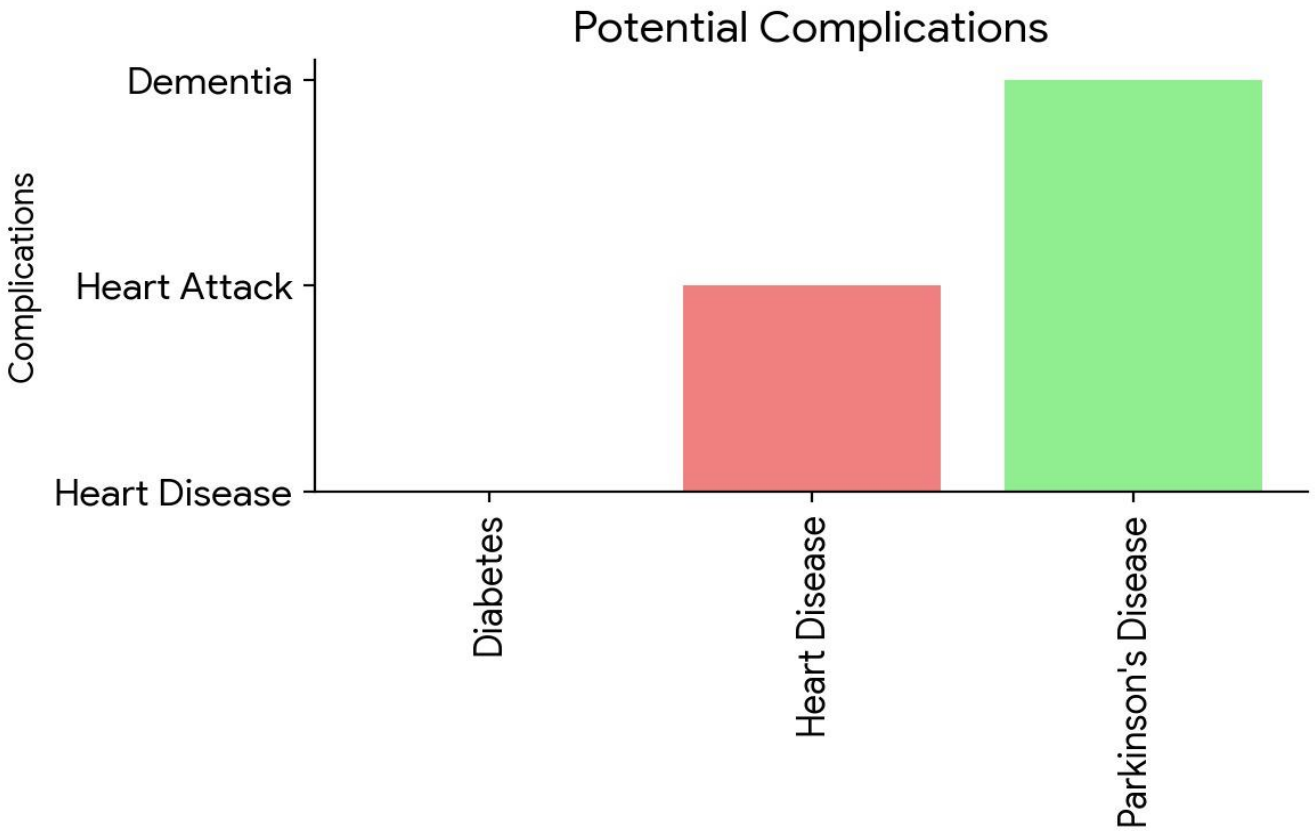



**fig. 5.5 Complication chart comparison between Diabetes, heart disease, parkinson disease**

## Description

- **Chart Focus:** It visually highlights the different complications that can arise from each disease.
- **Disease Breakdown:** Each disease (Diabetes, Heart Disease, Parkinson's Disease) has its own section in the chart.

- **Complication Details:** Each section likely lists the specific complications associated with that disease.
- **Comparison Potential:** By comparing sections, you might see how these diseases affect different bodily systems or have varying severity levels of complications

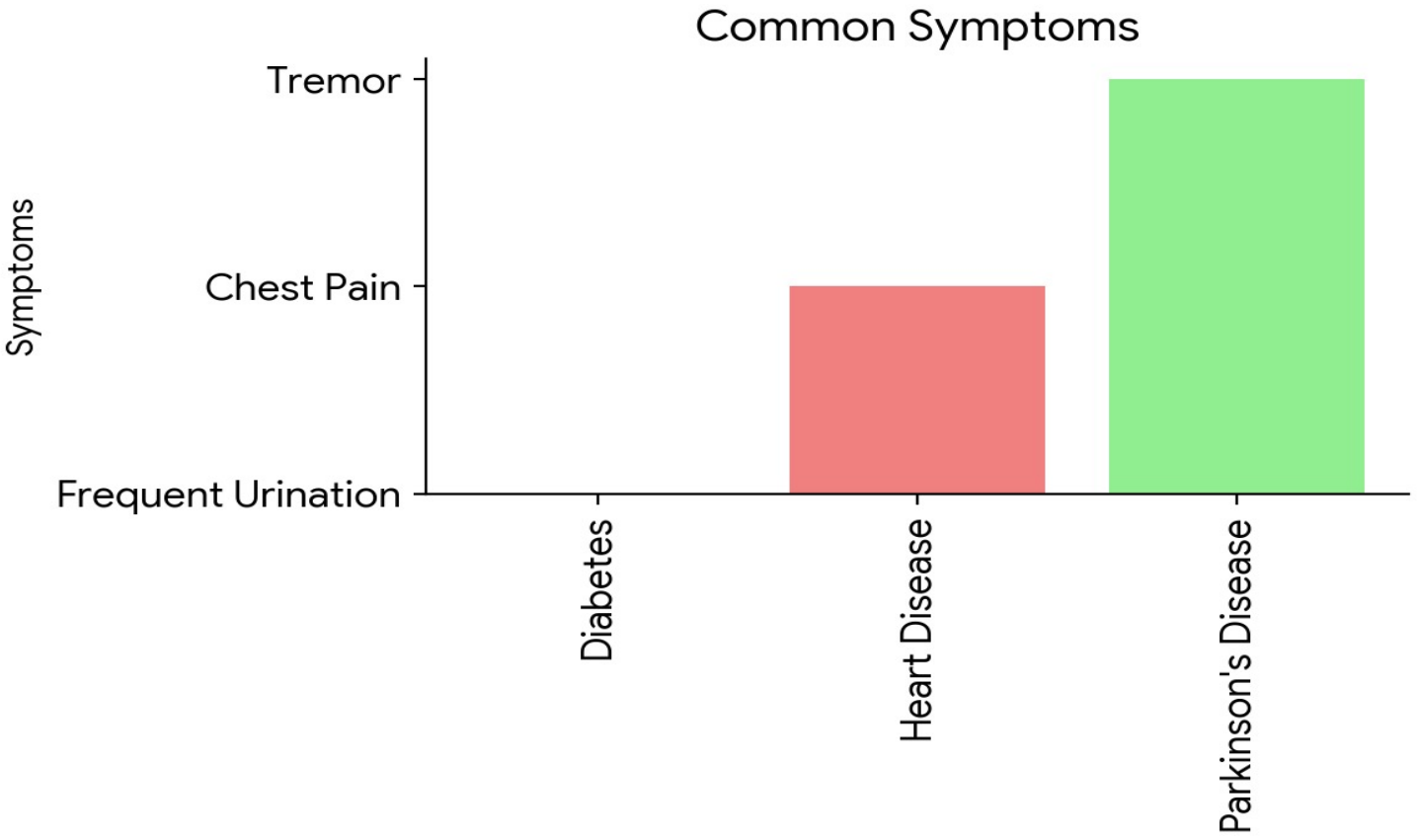



**fig. 5.6 Symptoms chart comparison between Diabetes, heart disease, parkinson disease**

## Description

- **Disease Focus:** It highlights three conditions: Diabetes, Heart Disease, and Parkinson's Disease.

- **Symptom Comparison:** Each disease would have its own section listing its specific symptoms.
- **Shared vs. Unique:** The chart might differentiate between overlapping symptoms (common across diseases) and those unique to each disease.

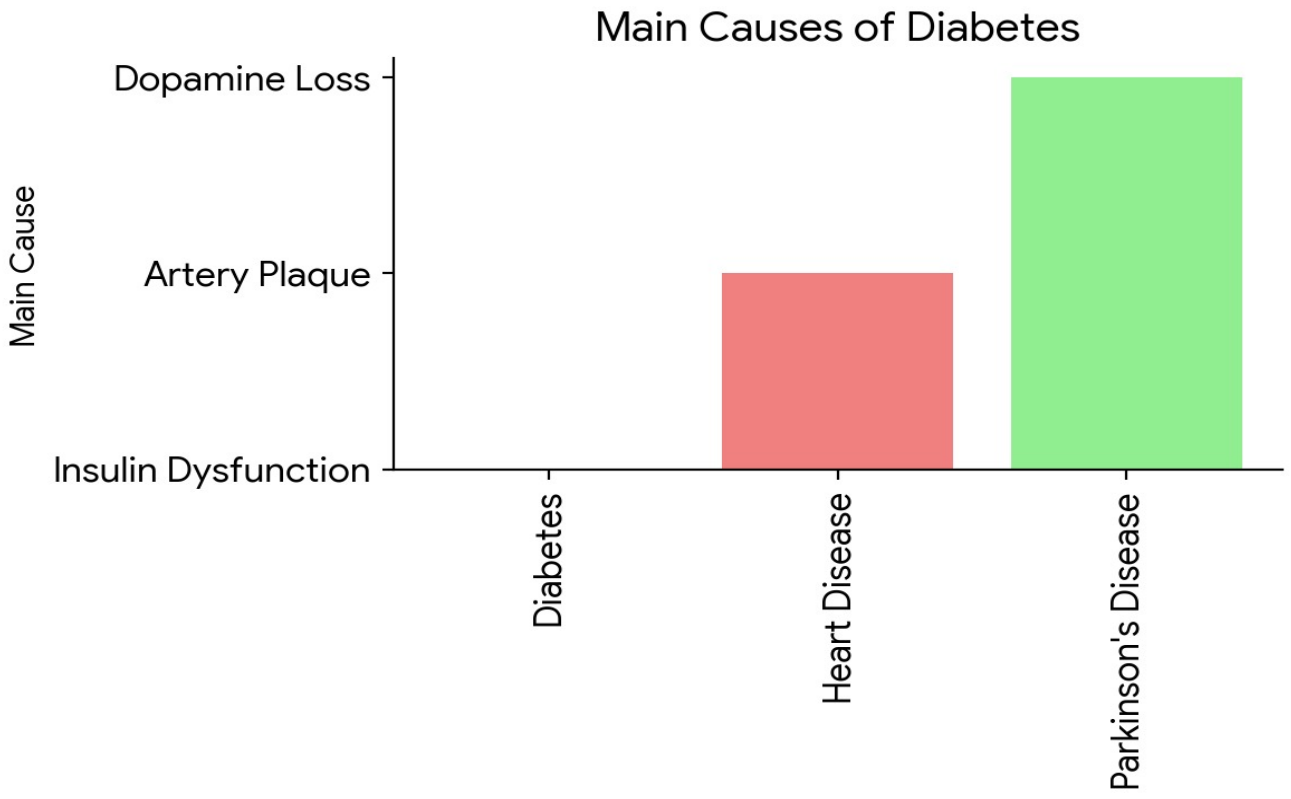



**fig. 5.7 Causes chart comparison between Diabetes, heart disease, parkinson disease**

## Description

- **Chart Format:** The chart likely has rows for each disease (Diabetes, Heart Disease, Parkinson's Disease) and columns for different categories of causes (e.g., Genetics, Lifestyle, Environmental).
- **Cause Comparison:** Each cell would indicate the strength or relevance of a specific cause for each disease.

- **Genetic Predisposition:** The chart might highlight a genetic component for all three diseases, but with varying degrees of influence.
- **Lifestyle Factors:** Unhealthy diet, lack of exercise, and smoking could be listed as risk factors for all three, but perhaps more strongly linked to heart disease and diabetes.

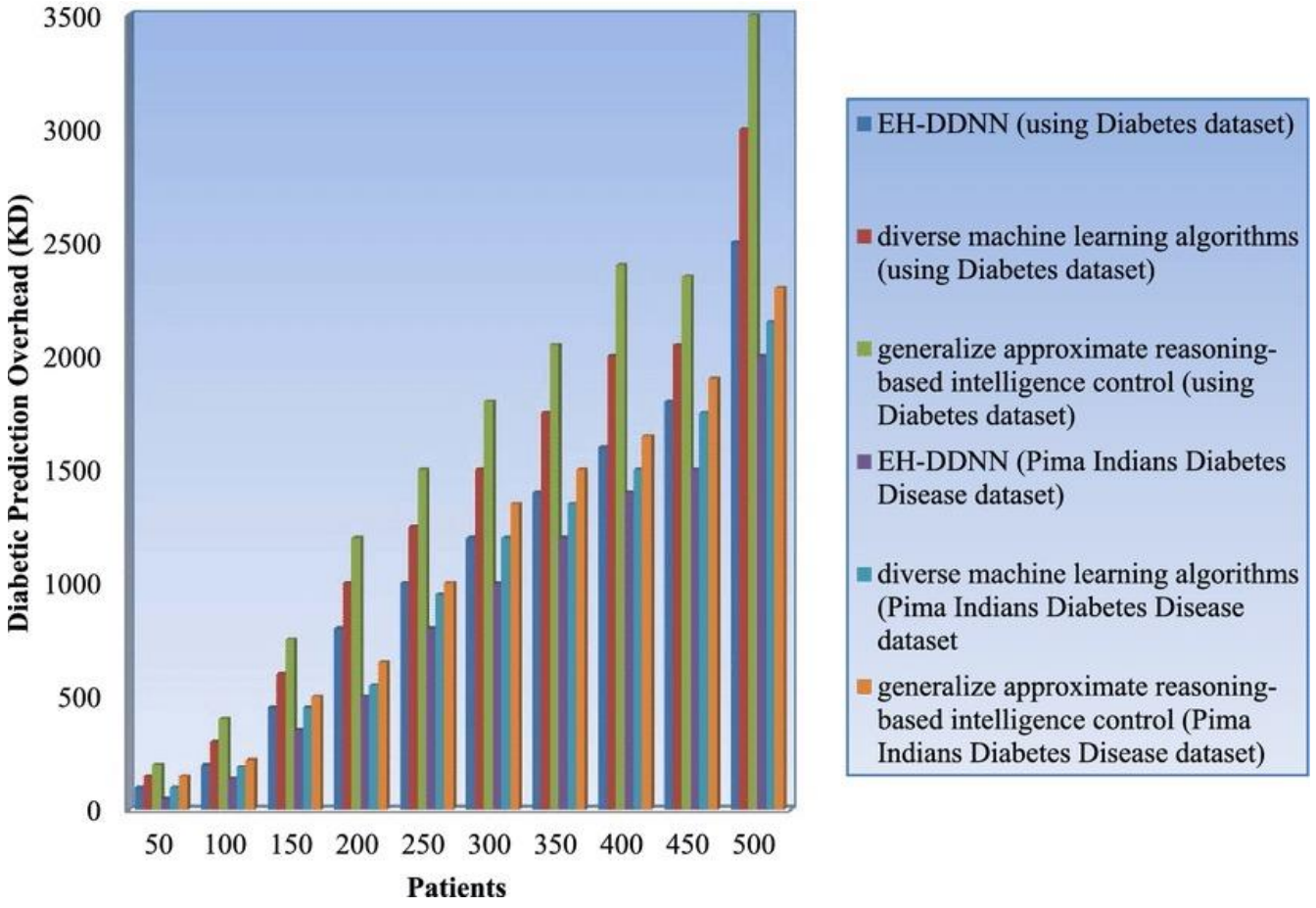



**fig. 5.8 Graph representation for diabetes**

## Description

- **Blood Glucose Levels:** A graph can show how blood glucose levels change over time, with time on the x-axis and glucose concentration on the y-axis. This can help visualize trends and identify potential issues.

- **Risk Factors:** A graph can connect different risk factors for diabetes, like obesity, age, or genetics, to illustrate how they interact and influence disease development.
- **Disease Progression:** A graph might depict the progression of diabetes complications, connecting factors like high blood pressure or neuropathy to potential consequences on organs like kidneys or eyes.

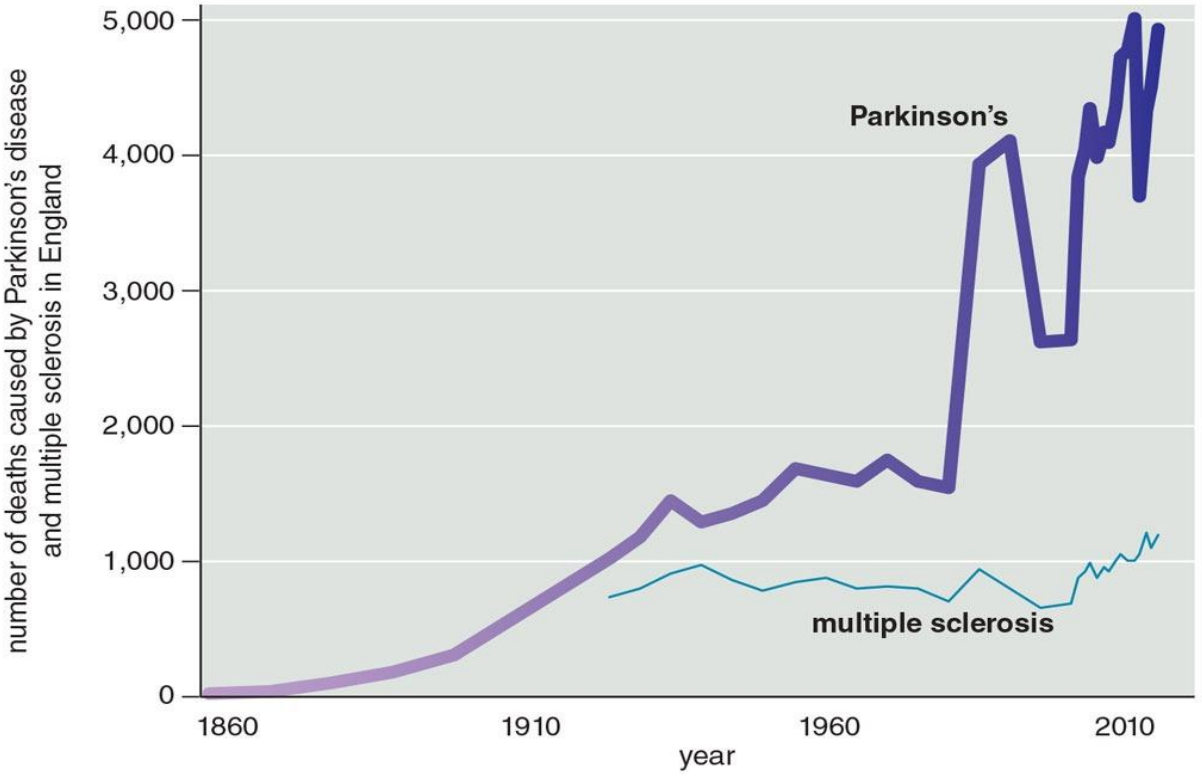



**fig. 5.9 Graph representation for Parkinson disease**

## Description

- **Nodes:** Represent entities like brain regions, genes, or symptoms associated with Parkinson's disease.
- **Edges:** Connect the nodes, indicating relationships between them. Stronger connections might suggest a more significant influence.

- **Analysis:** By analyzing the connections and their properties, researchers can explore potential disease mechanisms and identify relevant factors for diagnosis or treatment.

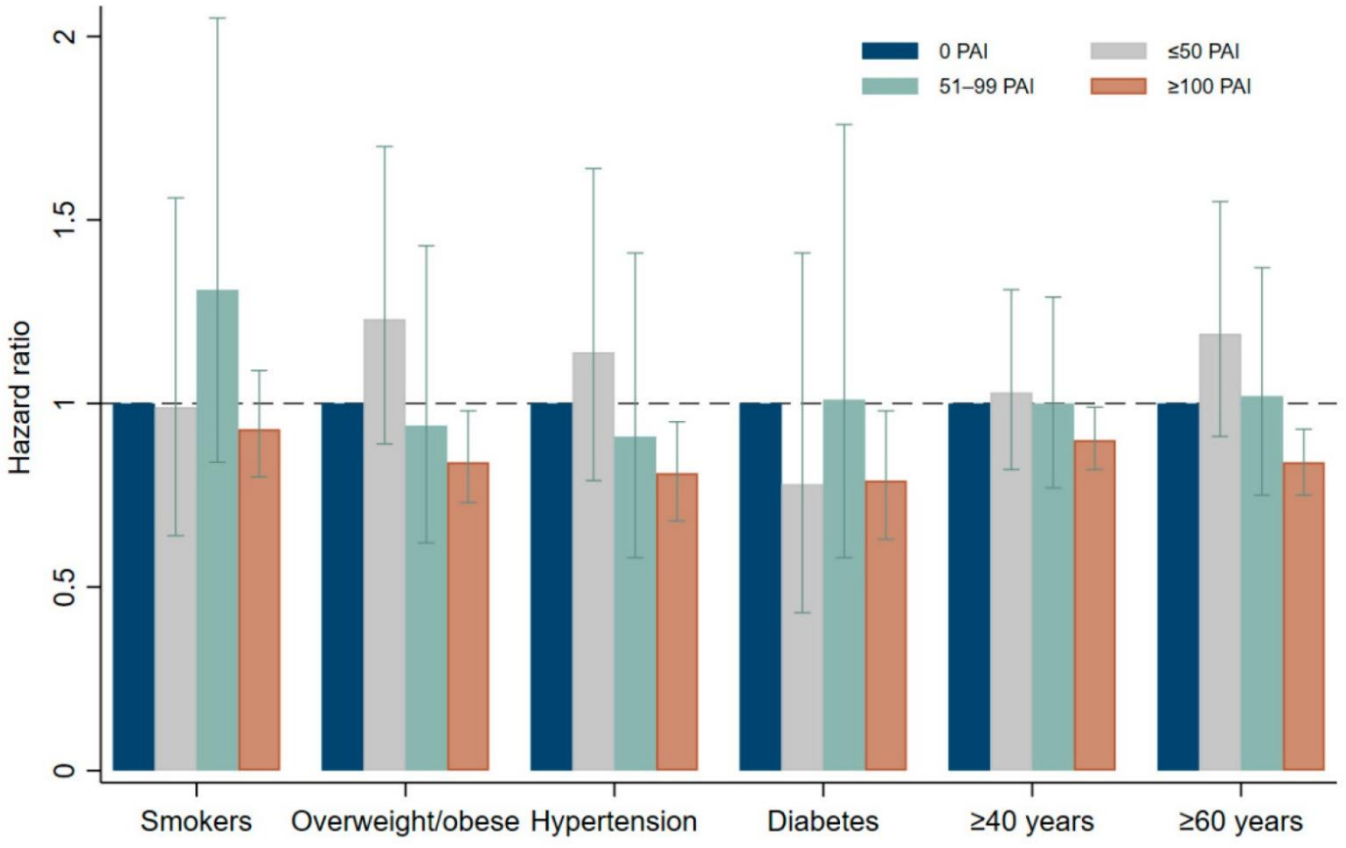



**fig. 5.10 Graph representation for heart disease**

## Description

- **Nodes:** Represent entities like patients, genes, or medical tests.
- **Edges:** Connect nodes, indicating relationships between them, like a patient having a specific gene mutation or a test result influencing another.
- **Edge properties:** Can encode additional information, such as the strength of a genetic association or the temporal order of tests.

# CHAPTER 6

# LITERATURE SURVEY

Review on Fuzzy and Neural Prediction Interval Modelling for Nonlinear Dynamical Systems
Oscar Cartagena;Sebastián Parra;Diego Muñoz-Carpintero;Luis G. Marín;Doris Sáez
IEEE Access
Year: 2021 | Volume: 9 | Journal Article | Publisher: IEEE

**Sources of Uncertainty in System MODELING**

In order to fully understand the information provided by Prediction Intervals, it is important to study what are the possible causes of uncertainty in dynamical system modeling, and on what phenomena they can originate.

In theory, predictive modeling assumes that all observations are generated by a data generating function f(x) combined with and additive noise

$y=f(x)+\epsilon$

where $\epsilon$ is a zero mean random variable called data noise that is responsible for introducing uncertainty modeling into predictive models in the form of **aleatory uncertainty**, whose origin can be traced to the exclusion of complicated variables from the model which cannot be determined with sufficient precision, or due to the presence of inherently stochastic processes in the observed system or data obtention procedure.

Based on this formulation, predictive models attempt to produce an estimate $\hat{f}(x)$ of the data generating function in order to calculate predictions of the expected value of the system. The model of $\hat{f}(x)$ is often referred to as crisp model. This procedure introduces an additional form of uncertainty, known as **epistemic uncertainty**, since the crisp model is only an approximation of the true data generating function. Assuming both types of uncertainty are independent, the total variance of observations can be expressed as

$\sigma^2_{total}=\sigma^2_{model}+\sigma^2_{data}$

where σ2model is attributed to epistemic uncertainty and σ2data is attributed to aleatory uncertainty. Since the factors contributing to epistemic uncertainty can vary greatly, some authors have proposed subsequent classifications for this term:

1. **Model misspecification:** Uncertainty determined by how close the estimate f^(x) can approximate the real data generating function f(x) under optimal parameter and data conditions
2. **Training data uncertainty:** Uncertainty over how representative the training data is with respect to the whole input distribution, and how sensitive the model can be to unseen samples

3. **Parameter uncertainty:** Uncertainty on the values of the model parameters due to local minima stagnation or premature termination.

It is important to note that since total uncertainty can come from many diverse sources, the expression can be highly complex and difficult to quantify, which is why interval modeling has been proposed as a solution to this problem.

Big Medical Data Decision-Making Intelligent System Exploiting Fuzzy Inference Logic for Prostate Cancer in Developing Countries

Kanghuai Liu;Zhigang Chen;Jia Wu;Yanlin Tan;Leilei Wang;Yeqing Yan;Heng Zhang;Junyao Long

IEEE Access

Year: 2019 | Volume: 7 | Journal Article | Publisher: IEEE

Introduction

Prostate cancer (PCa) has become the second commonest malignant tumor and the fifth leading cause of high morbidity and mortality in males, and it poses a rising public threat to human beings all over the world. To be specific, in 2012, more than 1.1 million males worldwide were diagnosed with prostate cancer. Additionally, the average morbidity rate of prostate cancer is approximately 11% over a male's lifetime, while the mortality rate from the disease is about 4% . In 2018, people suffering

from prostate cancer in Asia account for the half of 18.1 million new cancer cases all over the world. Meanwhile, the risk of prostate cancer has constantly risen to 13.5% which stands at the second highest among all cancers in male.

Some developing countries in Asia and Africa, such as China, India, and South Africa, may face the same challenge in social healthcare field that is a deep contradiction between huge population and scarce medical resource (undeveloped medical technology and insufficient public healthcare services). In particular, the repeatability and complexity of medical diagnosis program, a massive influx of multimodal medical data, and behindhand medical equipment will lead to a high misdiagnosis rate and relatively low diagnosis efficiency of medical staff eventually.

In Beijing, one of China's metropolis, there are over 20 million workers and 10 million children and the aged. However, in such a densely populated city, only no more than 3,000 healthcare personnel can provide healthcare service. Moreover, these healthcare personnel also must deal with hundreds of pathological reports from remote areas with behindhand medical system, because advanced medical devices and excellent physicians are centralized in big cities and first-class hospitals. What's more crucial is that heavy work and mental stress have already constituted a huge burden to those physicians, while thousands of patients are still waiting for malignant tumor diagnosis. Proportionately, in China, over 5000 patients must share one physician, and one general practitioner needs to take care of at least 70 prostate cancer patients a day. Over the whole diagnosis cycle, a mass of multimodal medical data associated with prostate cancer will be extracted, but only 30% of these statistics can be useful for the detection, diagnosis, and prognosis of the disease.

A New Collaborative Filtering Approach Based on Game Theory for Recommendation SystemsSelma Benkessirat;Narhimene Boustia;Rezoug Nachida

Journal of Web Engineering

Year: 2021 | Volume: 20, Issue: 2 | Journal Article | Publisher: River Publishers

Several recommendation systems were proposed in the literature. Some of these systems were proposed as general systems, which can be adapted in any domain of application. Other systems have been designed and incorporated in a several aspect of Web engineering. Indeed, RS Can improve extensively the quality of Web services for example to retain customers of a given brand, promote

tourism in such a country, improve the quality of distance education, etc. (response of RQ1). Consequently, lots of Web based applications were developed notably in e-commerce, e-learning, e-tourism, etc. (response of RQ2). These applications were developed using a set of techniques like Collaborative Filtering, Content-Based filtering, fuzzy logic and especially machine learning techniques such as clustering based approaches (response of RQ3). With this in mind, we classify the studied approaches into two categories, Clustering based approaches that exploit clustering and the approaches have been proposed as a solution for a specific domain. In the following, we present this two categories starting by Clustering based approaches. It is worth noting that each presented approach is discussed based on the exploited models and its validation (response of RQ4).

**Clustering Based Approaches**

Clustering based approaches take into account the similarity between users to more than one level. Clustering makes it possible to have similar user groups; it is a preselecting that takes into account the similarity between the users. After obtaining user clusters, a recommendation process is applied on each cluster independent of others. Many researchers integrated the clustering to the recommendation model.

In, authors have proposed clustering with regards to recommender system by employing the k-means algorithm as a pre-processing step in order to aid in neighbourhood formation. As a pre-selection step for neighbours, they employed the distance from the user to various cancroids. Experimental outcomes indicate that the recommended structure can considerably enhance the precision of predication and scalability issue. The datasets utilised for the experimentation are minor to evaluate the scalability problem.

In Linqi et al., an algorithm has been designed that separately explores the context space as well as the item space, and creates an algorithm that integrates clustering of the items with that of information aggregation pertaining to the context space. The proposed framework lacks a test and validation.

Son et al.have put forward an empirical study that integrates Dimension Reduction as well as User Clustering in Collaborative Filtering. A pre-processing phase is employed as the available algorithms. Post that, it employs traditional approaches to deal with these issues. The quality of the model has not been evaluated. An evaluation and comparison with conventional approaches would be necessary to

validate it. The results of processing time have shown that the performance of the developed system is significantly enhanced.

Zarzour et al.have developed a new collaborative filtering recommendation algorithm by considering the dimensionality reduction as well as clustering techniques. In order to reap benefits from the advantages pertaining to each algorithm, in this solution, the Singular Value Decomposition (SVD) and k-means algorithm were both used. The experimental results showed that the proposed method improved significantly the performance of the recommendations and remained the lowest values in the RMSE curve in the whole neighbors range.

The presented framework is based on Hierarchical Clustering. The clusters are formed by employing the Chameleon Hierarchical clustering algorithm. As per the experimental results, the put forward framework was seen to yield less error versus K-means based Recommender System. However, lower running time complexity was associated with K-means based RS versus the put forward method.

**A New Recommendation Approach Based on Probabilistic Soft Clustering Methods: A Scientific Documentation Case Study**

Remigio Hurtado;Jesús Bobadilla;Rodolfo Bojorque;Fernando Ortega;Xin Li

IEEE Access

Year: 2019 | Volume: 7 | Journal Article | Publisher: IEEE

Experiments Design

This section explains the experiments design: chosen datasets, soft clustering tested methods, quality measures, parameter values, etc. Experiments are performed using cross-validation. We compared recently published approaches to FCMwith our prediction approach *FCM-SP*. Our approach applied to BNMF yields similar quality results, since this model does not admit significant improvements.

The chosen datasets are the public *Movielens 1M* and the *SD4AI* ones. The soft clustering methods are *BNMF* [10] and our proposed method *FCM*. Finally, we test diverse clustering quality measures (*F-Partition*, *cohesion*, *separation* and *Xie and Beni index*), and the *MAE* prediction quality measure.

To run the designed experiments, we have chosen two open datasets: *Movielens 1M* and *SD4AI*. These two RS collaborative filtering datasets are very different, and this circumstance will help us to compare the performance of the clustering methods: *BNMF* and *FCM*. *Movielens* is a classical collaborative filtering dataset, containing votes casted by users to movies. *SD4AI* is a scientific documentation datamined dataset containing cardinalities of topics from each paper. Table 2 shows the main parameter values of both datasets: they have a similar number of ratings, but *SD4AI* holds much more papers and topics than *Movielens* users and movies does; consequently, *SD4AI* is sparser than *Movielens*. On the other hand, the ranges of votes/cardinalities of the tested datasets are radically different: to obtain accurate *SD4AI* predictions will be more difficult than getting it using *Movielens*.

γx,j=ϵ+i,j=ϵ−i,j=ax,j=Cj=α+∑{i|rx,i≠·}λx,i,jβ+∑{x|rx,i≠·}λx,i,j·r+x,iβ+∑{x|rx,i≠·}λx,i,j·r−x,iγx,j
γx,1+…+γx,K∑nx=1(ax,j)rx∑nx=1(ax,j)|rx,i≠·}
where: Cj is the centroid of cluster j and rx are the ratings (or cardinalities) of x
Px,i=ax,j=Cj=∑j=1…Kax,j·Cj,i1∑Kl=1(||rx−Cj||2||rx−Cl||2)2(m−1)∑nx=1(ax,j)mrx∑nx=1(ax,j)m
The quality measures we use in the experiments are:

- *F-Partition coefficient*: this parameter measures the amount of overlap between clusters. If F is 1 there is no membership sharing (extreme hard clustering). If F is 0 there is a total membership sharing between clusters (extreme soft clustering). Most real situations require a balance in the membership sharing between clusters.
- *Soft compactness (soft cohesion)*: it measures the distances from each item j (Xj ) to each cluster i (Vi ). Since we are using soft clustering methods, we weight each distance by using μi,j : the probability of item j to belong to the cluster i . The soft compactness can be formalized as:

  compactness=∑ci=1∑nj=1μ2i,j‖Vi−Xj‖2n

  where n is the number of items, and c is the number of clusters. Low soft compactness values are better, since it means that items belonging to each cluster are near.

- *Soft separation*: it measures the minimal distance between any pair of the cluster centroids. The soft separation can be formalized as:

dmin=mini,j‖Vi−Vj‖

where i and j are any pair of cluster centroids. High soft separation values are better, since it means that clusters are more separated.

- *Xie and Beni index (XB)*: it just makes the compactness divided by separation. XB provides a unified value for the above clustering quality measures. Its drawback is that it loses the details of both compactness and separation values.

- *Mean Absolute Error (MAE)*: this is not a clustering quality measure; it is a collaborative filtering prediction quality measure. The MAE can be formalized as:

MAE=1|{ru,i|ru,i≠·}|∑u∈U∑i∈I|pu,i−ru,i|

where ru,i is the user u rating to the item i . pu,i is the prediction of ru,i , and {ru,i|ru,i≠·} is the set of casted ratings. Low MAE values are better, since it means that prediction errors are lower.

**Learning Object Recommendations for Teachers Based On Elicited ICT Competence Profiles**

Stylianos Sergis;Demetrios G. Sampson

IEEE Transactions on Learning Technologies

Year: 2016 | Volume: 9, Issue: 1 | Journal Article | Publisher: IEEE

**User Profiling**

User profiling is a technique that has been widely applied in a range of software applications, including adaptive web systems and recommender systems. It involves gathering data from the users in order to create a profile for each of them, depicting their unique attributes in the context of the system's application. The latter could involve, for example, movie genre preferences in a movie recommender system, or learner preferences in an educational RS.

The process of creating individual user profiles in software applications is essential because each user has his/her own characteristics and needs. Therefore, capturing these user attributes is necessary for

enabling the provision of personalized services. A user profile, therefore, presents the system's full interpretation of the users' preferences and personal characteristics.

In many cases, user profiles are not provided by the users themselves but they are being elicited automatically through the identification, collection and processing of relevant user actions' data . The reason for this is that users are usually either unwilling to provide such information or when they do, the validity and completeness of the provided data cannot be ensured. Therefore, the elicitation process is usually based on the users' relevance feedback data.

Relevance feedback data can be attained in two ways, namely explicit feedback and implicit feedback. The former requires users to perform specific actions that will inform and update their profile attributes, e.g., assign ratings to items or download items. This approach provides a set of benefits for the hosting RS system, such as increased development simplicity and enhanced accuracy in the profile update process. Users' implicit feedback refers to mechanisms that monitor the users' interaction with the system in an unobtrusive manner. Such approaches have been developed in order to completely detach the user from the explicit feedback providing process and maximize the amount of data that are being harvested by the system. Examples of user actions that are being monitored for profiling purposes include browsing time in each item and type of items accessed, uploaded or ignored.

User Profiling in Teacher-Oriented Recommender Systems

In the context of TeL, the majority of the implemented RS targets the learners and aims to provide them with personalized learning material and sequences of learning activities towards specific educational goal attainment . The main *learner attributes* that are being used in such processes include their prior knowledge, learning preferences/styles, individual goals and other cognitive characteristics. Subsequently, user profiling approaches have been primarily focused on accommodating the attributes of the learners .

However, despite the fact that learners are indeed the main focus of learning processes, other actors are also important for the successful implementation of effective learning procedures. More specifically, ***teachers*** also play a vital part in the educational processes and, amongst other reasons, their ICT competences and personal attitudes towards ICT use can greatly affect the level and the

quality of their technology-supported teaching practice. Therefore, systematic accommodation of teachers' professional characteristics should be an important design consideration for educational RS. Within this context, our previous work included a literature review of existing teacher-oriented educational RS. The 22 identified approaches are summarized in Table 1. Table 1 also contains information on the types of relevance feedback data that are being harvested. These relevance feedback data include Explicit and Implicit data. The former include Social data (e.g., ratings and bookmarks) and user Demographic data, while the latter include, for example number of views of LO. Moreover, Table 1 presents information on whether these data are being utilized only for ad-hoc similarity calculations (i.e., they do not incorporate profile creation) or if they are being automatically processed and exploited to create dynamic and adaptive user profiles for providing personalized recommendations (i.e., they do incorporate profile creation).

**Merged Ontology and SVM-Based Information Extraction and Recommendation System for Social Robots**

Farman Ali;Daehan Kwak;Pervez Khan;Shaker Hassan A. Ei-Sappagh;S. M. Riazul Islam;Daeyoung Park;Kyung-Sup Kwak

IEEE Access

Year: 2017 | Volume: 5 | Journal Article | Publisher: IEEE

The use of social robots and IE from the Internet is a hot topic in the field of recommendation and information engineering research. The main advantage of social robots is to help humans in daily life. At present, people are employing social robots in public for different purposes, including education, recommendations, and healthcare activities. For recommendations, the social robot is unable to provide precise information because of the system limitations. Most of these systems are based on predefined words and sentences. However, based on these predefined words and sentences, a social robot cannot extract efficient information from the hazy environment of the Internet. Therefore, extensive methodological work in text mining is needed to automatically get the FTQ from the social robot, re-treat the FTQ, and then retrieve the information based on the query. A variety of research has been done on this issue, and several ideas on text mining for social robots have been proposed.

Unified communications and rich web interactions are employed to help a Neel robot connect with humans in an interpersonal manner in a shopping mall. The rich web interaction enables the robot to make links with users and help customers to get deals and offers. However, this system may not

recommend a precise item because it trusts the data updated by sellers. Therefore, this system needs preference data. A coupon-giving robot system was developed for advertising in a shopping mall. In this system, two tests show the efficiency of a recommendation: varying conversation schemas, and the existence of a robot. The presence of a robot strongly attracts customers. However, there is a limitation to rules for recommendations, and a category of words is used for speech recognition, which affects the accuracy of results. A text mining–based recommendation system was developed for human–robot interaction. In this system, the robot communicates with the human to get the oral query, and then obtains the information employed in an external corpora. This system recommends a movie according to a user's oral query. The idea of robot rules–based recommendation for retail shops was proposed. This system tracks the customer's location and recommends precise products that are near them. The system needs semantic rules to recommend products. Human–robot interaction was examined by using various social activities in a hotel. In this system, one robot detects and greets the customers, and another converses with body gestures, so the guest obtains information by talking. A cloud robotics service is employed in a smart city for emergency management operations. This system uses small unmanned aerial vehicles (UAVs) as agents connected to the network. This network infrastructure allows the UAVs to benefit from storage resources and to control open data from common knowledge. A NAO robot and a smart pen are employed to improve social communications with dementia patients. The NAO robot constantly monitors patient activity and assists the patient in daily life situations. In this system, patients can communicate with the NAO robot by using speech functionality. The NAO robot and multiple cameras are also employed for colored object recognition. This system allows the NAO robot to receive oral commands from users to find a needed colored object. Fuzzy logic is employed to recognize the color based on the user's perception and multiple cameras to improve the quality of the recognitions.

At present, the increase in Internet data makes the recommendation debate more challenging. Most recommendation robot systems are based on predefined words and sentences, which may not allow the robot to extract precise information. Therefore, a full-text query can overcome this issue. The FTQ employs an existing search engine to obtain data about a particular topic or item. However, search engines use a keyword-matching mechanism, and are incapable of extracting the aim of the query from data on their servers. To overcome this problem, an intelligent effort is needed to re-treat the user query and to obtain the required information from the intensively blurred environment of the Internet. A full-text search query is employed to extract information from the Internet . This system

showed that an FTQ search performs well when keywords of the query are used in the extracted documents. Its primary concern is precision, not recall. However, there are some limitations in the system; the existing search engines employ keyword-matching mechanisms. Therefore, this approach is inadequate at extracting appropriate data from the heterogeneous sources of Internet information. Currently, sentiment analysis–based recommendation has become a hot topic in research. There are two approaches to sentiment analysis: sentiment classification and feature-level sentiment analysis. The text is categorized (positive, neutral, and negative) in sentiment classification approaches, and the information is intended for manually characterizing the sentiment words . In feature-level sentiment analysis, the features are described to extract sentiment words from text . The idea of summarization and sentiment analysis was presented , which defines the feature sentiment (negative or positive) of a product by employing a lexicon-based method.

Recently, an ontology has been employed in the area of recommendations, IE, and sentiment analysis. IE is used to transfer natural language text into structured information (Daniel et al., 2008). This transformation is obtained by identifying relevant concepts, relationships, and instances. However, natural language has ambiguity (single words can have many meanings). Therefore, the process of IE is a difficult task. Ontology-based IE overcomes this difficulty by organizing the domain knowledge through an ontology. An ontology is a shared conceptualization of a specific domain through concepts, instances, and relationships, which is in a human-understandable and machine-readable format. A regular ontology is applied to find features in movie reviews . This ontology is suitable for the extraction of a planned set of data. Nevertheless, Internet data are imprecise in structure. Consequently, a regular ontology is inadequate for describing the fuzzy terms of features (e.g., city {clean, average, and dusty}). A regular ontology with fuzzy logic works remarkably well when the input is uncertain. Opinion categorization of online item reviews was suggested to measure the linguistic hedges on sentiment labels . The system automatically extracts opinion phrases from reviews and categorizes the reviews in terms of positive and negative. Moreover, the sentiment scores are stored in linguistic variables and presented in table format. An ontology is employed to store all variables with an opinion score, and to provide a knowledgebase platform for the classification of feature polarity. Opinion mining based on an ontology was suggested to categorize and examine online reviews . The system shows that parts of speech can be different, which leads to ambiguous analysis, and decreases the accuracy of the review classifier. The system employs an SVM as an opinion analyzer to compute the precise measure of words for sentiment analysis.

Most of the existing research has its limitations in recommendations, IE, and sentiment analysis. Mostly, the recommendation and IE systems are based on predefined words and a regular ontology, respectively. It is a fact that predefined words–based systems are unable to recommend the correct item, and a regular ontology cannot extract the anticipated result from a blurred resource of data. To the best of our knowledge, the proposed merged ontology and SVM-based recommendation and IE is a first effort to automatically retrieve information and extract the meaning of the data for disabled users. This system can extract information related to hotels, the city, and diabetes drugs. In the proposed system, a NAO robot communicates with disabled user to get the oral query; the oral query is then mined to extract the disabled user's needs, and it is then converted to an appropriate form for an information search. Additionally, the proposed merged ontology provides fuzzy logic and semantic web rule language (SWRL)-based semantic knowledge for feature polarity computation. This merged ontology contains a medical ontology, a city ontology, and a hotel ontology, which easily provide information for recommendations.

**A Conscious Cross-Breed Recommendation Approach Confining Cold-Start in Electronic Commerce Systems**

S. Gopal Krishna Patro;Brojo Kishore Mishra;Sanjaya Kumar Panda;Amrutashree Hota;Raghvendra Kumar;Shiyang Lyu;David Taniar

IEEE Access

Year: 2023 | Volume: 11 | Journal Article | Publisher: IEEE

The memory-based approaches proposed by Guo et al. work openly with recorded interaction values and respond to nearest neighbour searches to find similar users. Meanwhile, this approach does not require any model assumption. For example, from the user of interest, the closest user is identified, and among these neighbours, the most famous items are suggested [23]. Li et al. proposed an effective recommendation mechanism based on the transactions in online business and a victorious e-commerce system. The proposed recommender system was able to generate recommendations with personalized products by taking into account social relationships, recommendation trust and preference similarity, although people in reality were influenced by the suggestions and opinions of people with similar interests, shopping interests, as well as close friends in reality [24].

Moreover, Shinde et al. reported the personalized recommendation technique for reducing the overload of information issues on the World Wide Web. They described a novel algorithm, such ascending-bunching-based clustering (CBBC), used to implement the centring-bunching-based clustering hybrid personalized recommender system (CBBCHPRS). This process helped in getting good quality and more effectiveness of the recommender mechanism for active users alleviating the issues like Sparsity, cold-start and first-rater [25]. Meanwhile, Nilashi et al. developed a novel hybrid mechanism of a recommender system based on collaborative filtering (CF) to get brilliant accuracy. This theory and mechanism were able to overcome the issue of Sparsity as well as scalability with improved efficiency. This mechanism was well-established by ontology and dimensionality reduction technology. Two real-world datasets were used in the study to evaluate the proposed model, which demonstrated improved effectiveness in addressing scalability and sparsity issues in CF [14].

The research contributions for the proposed CSSHRS procedure have been described as follows:

1. The customer might not contribute any kind of input in regard to every one of the things accessible on the stage. This absence of client appraisals on a specific thing lessens the proposal's precision. Thus, the inaccessible appraisals are at first anticipated before the genuine cycle.

2. The methodology of CSO (Cuckoo Search Optimization) is utilized in the bunching process, which assists with deciding an underlying focus of clustering of k-means and subsequently improves the accuracy and exactness of the framework of recommendation.

3. Also, LDA and PCA break down higher-request information to bring down aspects that can speed up processing. At long last, the procurement of information by means of manual power has been replaced by using the ANFIS framework. The preparatory interaction used in the proposed model limits the error, making the expectation fast and accurate without relying on individual specialists. Meanwhile, this approach allows for flexibility in the model rather than being completely rigid.

Sentimental Analysis Using Capsule Network with Gravitational Search Algorithm

V. Diviya Prabha;R. Rathipriya
Journal of Web Engineering
Year: 2020 | Volume: 19, Issue: 5-6 | Journal Article | Publisher: River

To overcome the limitations nature-inspired optimization are investigated for better sentiment analysis. However, the swarm intelligence takes inspiration from nature algorithm like Gravitational Search Algorithm (GSA) (Prabha and Rathipriya, 2013) is used to search space in a minimal architecture. It works on the concept of force and mass (Rashedi and Nezamabadipour, 2009) every words in the text attracts the other related words here particle act as agents. Search processes of each word are optimized to find the best capsule to classify the text. It also optimizes the values to find the best solution for identifying sentiments. Methods used to categorize tweets are discussed (Gohil and Vuik, 2018) in this methodology. Capsule Network (CN) (Kim and Jang, 2020) improves the performance in text analysis. It is an efficient method to understand text data and converting word to vector.

The contribution in this paper can be précised as follows: Design architecture of CN-GSA used improve the classification accuracy of both long and short twitter post about COVID-19, It helps to classify tweets as five classes such as strong positive, positive, strong negative, negative and neutral, Comparing the efficiency of the proposed model with machine learning and other deep learning model, The proposed model improves in terms of precision, recall, F-measure and accuracy, The model improves the accuracy for different kind of sentiment classification datasets.

In this work, the proposed Capsule Network (CN) with GSA is used to identify important key terms from sentence and optimize from large space text to minimized this will consequently improve the classification performance. This paper is organized as follow. The detail descriptions of proposed model are discussed in Section 3. Comparative analysis of CN with other models is represented in section 4. The results of experiments show that CN outperforms well and good compared with other methods. Finally, Section 5 concludes the paper.

**ScenarioSA: A Dyadic Conversational Database for Interactive Sentiment AnalysisYazhou Zhang;Zhipeng Zhao;Panpan Wang;Xiang Li;Lu Rong;Dawei Song**

**IEEE Access**

**Year: 2020 | Volume: 8 | Journal Article | Publisher: IEEEData Collection & Pre-Processing**

Our goal is to construct a large scale sentiment dataset to support the interactive sentiment analysis task. First, we crawl over 3,000 multi-turn English conversations from several websites that support online communication.[2] The conversations are collected in the various daily life contexts and cover a wide range of topics, such as shopping, work, travel, and food. More details of the topics will be introduced in Sec. IV-A. Each conversation is human written and thus is more formal than the transcribed text from a spoken corpus such as the First-encounter dialogue.

Since each conversation revolves around a certain topic, it usually ends after a reasonable number of turns (less than 25 turns in our ScenarioSA). The crawled conversations are clearly distinguishable from other dialogue datasets such as Cornell Movie-Dialogues Corpus and The NPS Chat Corpus.

All the conversations are then pre-processed. Some of the crawled conversations involve three or more participants. We think the conversation among multiple speakers will exacerbate the jumpings in logic of each speaker. In this work, we prefer studying the interactions between two speakers, and thus discard those involving three or more speakers. Further, for sake of privacy protection, we replace the first speaker's name with **A**, the second with **B**, and replace others' names mentioned in the conversation with **NAME**. We also correct the spelling mistakes automatically, and check if each conversation is composed of illegible characters.

After pre-processing, the ScenarioSA dataset contains 2,214 multi-turn conversations, altogether 24,072 utterances and 228,047 word occurences. The average speaker turns and average number of words per conversation is about 6 (turns) and 103 (words), respectively. The detailed statistics

Aspect-Based Sentiment Analysis Using a Hybridized Approach Based on CNN and GAAdnan Ishaq;Sohail Asghar;Saira Andleeb Gillani

IEEE Access

The structure of CNN can be seen, this consists of the separate layer stack which converts the volume of the input into target output by differential function. The constituent layers of CNN are

1. Convolution layer
2. Max-policy layer
3. Rely layer
4. Back propagation layer

The intention is to use sufficient filters (in this case, 128) to catch enough features in a particular sentence. In classification of images, different filters integrate different attributes for example edges, color density at various spots, transformation of colors etc. The problem of text classification stretches the similar idea to catch features for example “like” mean positive rather than similarity, “very much” communicate the degree of features utilizing filter with size

The fundamental aim is to provide sufficient range of features to grab all probable descriptor of the text. Maxpool shall have the highest output value of the vector on applying the filter. It picks a strongest expression aspect in the extracted function from the output and nothing to do with the length of the word. Every sample is represented as n*1 where n represents the length of dimension filter. This filter is practice as a drafting window, for instants 3*1 filter on a sentence. I like the car very much! would yield (I like this, this car very, car very much, very much!) sentence splits up to equal length ahead of embedding hence all the filter does not result in the identical dimensions outcome. The area size (2, 3, 4) is identical to 2, 3, 4 – G word and the first filter in this trigram will give different weight esteems to different words in trigrams. It suggest that higher weight are allotted to the first index (0-index) and lower to the second hence 128 filter will allocate respective weight that will be trained to optional weight value after sometimes for precise forecasting. For a particular sentence, a sentimental label is formed by scoring a sentence, by evaluating the order of words in a sentence as input and by passing through layers that extricate its feature with a high degree of complexity. Extraction of the features can be executed on lexical and character level. The

distinctiveness of network design is to incorporate two convolutionary layers that allow to handle any size of phrases and words.

**News Recommendation Systems in the Era of Information Overload**

Shuaishuai Feng;Junyan Meng;Jiaxing Zhang

Journal of Web Engineering

Year: 2021 | Volume: 20, Issue: 2 | Journal Article | Publisher: River Publishers

A recommendation system is developed by experts on the basis of computing technology and statistics, combining data, algorithms, and computers to create mechanisms that relate users with personalized resources that can reflect the user's consumer behaviors and information intake. To date, recommendation systems have developed a relatively complete methodology. The most common algorithm categories are collaborative filtering recommendations, content-based recommendations, association rule-based recommendations, utility-based recommendations, and knowledge-based recommendations.

First, we talk about collaborative filtering recommendations. These algorithms can be categorized into User-Based Collaborative Filtering and Item-Based Collaborative Filtering. User-Based CF finds other users that has a high similarity with the target user, i.e., similar users, and uses the preferences of those similar users to predict the preferences of the target user. Item-Based CF uses the ratings that users give an item or a message, analyzes the similarity between different items and messages, and recommends the items that has a high similarity with the items that the target user likes. The basics of these algorithms include association algorithms, categorization algorithms, clustering algorithms, regression algorithms, matrix decomposition, graph models, word connotation models, and neural networks. Next, there are content-based recommendations. These algorithms do not need to utilize the item ratings of users, but instead use the browsing history of a user to predict what the user might have never seen but might like. Third, there are association-based recommendations. These algorithms smartly utilize the ability to find association in big data. By using the association rule and digging through datasets, the relationship between different items during their sales and usage can be discovered and be used to predict the users' needs. Fourth, there are utility-based recommendations. The core of this algorithm is to create an utility function for each user that has the item characteristics and customer satisfaction as input. After calculating the different values of utility given the different inputs, the item with the highest utility is recommended to the user. Fifth, there is knowledge-based

recommendations. This algorithm uses the knowledge structure in the user data that can support inference (such as regularized user search history), creating a knowledge base of how an item can satisfy a particular user, and using that knowledge to make a prediction [4]. In these common categories of recommendation algorithms, collaborative filtering is the main algorithm basis of news recommendation system. We talk about the principles and processes of collaborative filtering below.

**Design and Implementation of Writing Based on Hybrid Recommendation Systems Recommendation**

Langcai Cao;Biyang Ma;Ya Zhou;Bilian Chen

IEEE Access

Year: 2018 | Volume: 6 | Journal Article | Publisher: IEEE

we propose a hybrid recommendation method for two existing problems mentioned before, that is, one is to introduce users' background similarity to calculate new users' neighbor users and fill them by using an average score, the other one is to reduce the adverse effect of data sparsity on recommendation accuracy.

Different users have different background information. This information has an important influence on users' personal preferences, and the recommendation system should consider this factor. Therefore, we propose a method that considering candidates' background information when designing a recommendation system. Background information includes its own level, test score target, and test time. For example, when using collaborative filtering to make recommendations, we should consider the users' own comprehensive information. Since most website users are registered, we can collect personal information by filling out a questionnaire during registration. For example, after completing the user registration process in writing practice websites, the web page directly jumps to the improvement page of personal information. On this page, the information of the candidates is collected, including test subjects, IELTS/TOEFL, expected test scores, last test scores, and test time. Considering the actual situation of TOEFL/IELTS, users with similar information can also have high similarities in the demand for test resources. Therefore, the personal information data that users filled in during registration can be used to solve the cold start problem of new users in the IBCF algorithm, thereby supplementing the algorithm.

By calculating the similarity of background information, users with the highest similarity of background information are found for new users as neighbor users. The recommendation is implemented for new users based on the related information of neighbor users, thereby solving the cold start problem.

The purpose of this writing practice website is to provide an exercise platform for candidates taking IELTS or TOEFL. This website is different from traditional e-commerce websites. Users do not need to evaluate the questions, and no questions of interest exist. Due to the nature of IELTS/TOEFL, the official examination authorities publish the average score of Chinese candidates each year. We can use users' score on a certain test item as the data of the user-item matrix. For the default value, we use the official average. This approach is a convincing solution. Through this analysis, the algorithm is still based on collaborative filtering.

**Deep Learning-Based Recommendation Systems: Systematic Review and Classification**

Caiwen Li;Iskandar Ishak;Hamidah Ibrahim;Maslina Zolkepli;Fatimah Sidi;Caili Li

IEEE Access

Year: 2023 | Volume: 11 | Journal Article | Publisher: IEEE

Using specific search terms, our findings are categorized based on various aspects, such as techniques, domains, and types of recommendation systems.

Our goal is to systematically review the current state of deep-learning-based recommendation systems and highlight key findings in each category. To achieve this, we follow a systematic literature review (SLR) process similar to the one used in the article by Alabadla et al. We adhere to the guidelines from Kitchenham et al. and follow the best practices recommended in Wang et al.'s study [39] to ensure the rigor and validity of our review.

We include research questions and sub-questions, search strategy, literature search, screening process, quality assessment, data extraction, and data synthesis, all of which we describe in detail. Our systematic review is conducted using specific search terms, and we assess the quality of the studies included in the review using appropriate quality assessment tools.

Our review process helps identify research gaps and provides a better understanding of the key concepts and variables associated with deep-learning-based recommendation systems, contributing to future research in this area.

We use these research questions and sub-questions to guide our literature review and data collection process. The search terms relevant to our study are identified based on these questions. They retrieve relevant literature from various sources, such as online databases and search engines.

This systematic review poses important research questions on deep learning-based recommendation systems. RQ1 identifies popular deep learning techniques, such as autoencoder, CNN, RNN, and MLP, along with their evaluation metrics, such as accuracy, precision, and recall. It also highlights the datasets used in previous research, such as MovieLens, Amazon, and Yelp, providing researchers with insights into potential gaps and opportunities for future research. RQ2 categorizes the different applications of deep learning techniques, providing valuable guidance on how to apply these techniques in various domains, such as e-commerce, social networks, and e-learning.

News Recommendation Systems in the Era of Information OverloadShuaishuai Feng;Junyan Meng;Jiaxing Zhang Journal of Web Engineering

Year: 2021 | Volume: 20, Issue: 2 | Journal Article | Publisher: River Publishers
As the scale of recommendation systems increases and there are tens of millions of users and news content, the possibility of some pieces of chosen news overlapping between two users becomes very small. If we look at data sparsity using a ratio of the existing relations between a user and some pieces of news to all possible existing relations, we can observe that as the number of news content grows exponentially, data becomes more and more sparse, which will significantly increases the computation complexity of the system. The root of this problem cannot be entirely solved, but there are some ways to reach a middle ground. If it is possible to use a kind of expanding algorithm, changing the original order-one relations (how similar are two users or the news content that read) to relations of order-two or higher (given that relations and similarity itself is expandable) , then some default parameters can be used and increase the resolution of similarities. The bigger the scale of data, the sparser it usually is. There are some sparse data algorithms that are believed to be very helpful in the future (e.g. expanding , iterative optimization , similarity transference etc.)

Second, there is the problem of cold starts. For a new user that has just joined a system, there is no effective behavioral data that can be used as reference and it is therefore hard for the system to give the user a precise recommendation. On the flip side, some news content has only been seen by users a few times and it is therefore difficult to recommend this piece of news to users. One solution is to use the content as an assistive recommendation, and another is to gather some data such as age, city, education level, gender, and occupation upon signup or via a questionnaire. Recently, tagging systems that have been used in many places are also a possible solution , because tags can been thought of as the essence of content, and they simultaneously reflect the personal preferences of users. Two users can watch the same news but be interested in different parts of the same piece. Of course, the usage of tags can only improve recommendations given to users with minimal behavioral data but does not help users who are from a completely cold start, because no tags are associated with these users.

**Rating Prediction Based on Merge-CNN and Concise Attention Review Mining**

Yun-Cheng Chou;Hsing-Yu Chen;Duen-Ren Liu;Der-Shiuan Chang

IEEE Access

Year: 2020 | Volume: 8 | Journal Article | Publisher: IEEE

Deep learning methods, such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), have been used successfully in many fields, such as computer vision , natural language processing (NLP) and document classification, and so on. Deep learning is successfully applied in text analysis to extract opinions or sentiments from online reviews. Extracting users' opinions needs several NLP steps to identify subjectivity in the given text .

Extracting sentiment is to assign the polarity (e.g. positive, neutral, and negative) to the sentences or text by using deep learning methods based on the sentiment analysis. CNNs and RNNs learn data from multiple deep layers of modules. Instead of image pixels, the inputs can be sentences or documents represented as a matrix. CNNs have the ability to extract important n-gram features from sentences to generate the latent semantic representations for NLP tasks, such as sentence, sentiment, and subjectivity classifications .

RNNs include the notion of time in neural network models and can carry previous information to the current neural state. Unlike CNN characterized by its ability to extract regional features, RNNs are characterized by their ability to model units in a sequence. User review content usually represents the user's most realistic emotional response. Hochreiter and Schmidhuber designed the long-short-term-memory units (LSTM) to optimize neural units in general RNN. Moreover, a hierarchical structure was proposed for sentiment classification, in which a document is composed of sentences, and sentences are composed of words. Moreover, gated recurrent units (GRUs) were introduced. GRUs have fewer parameters than LSTM and do not have an output gate. A bi-direction GRU can be used to extract both forward and backward features from sentence representations for generating the document representations and sentiment classification.

The attention mechanism is an interesting design that simulates the habit of attention in the human brain; it helps to focus on important information and improve the effectiveness of information processing. Different sentences and words make different contributions in representing the meaning of a document. A hierarchical attention network based on attention mechanism and bidirectional GRU (Bi-GRU) was proposed for text categorization . Their proposed hierarchical architecture shows the different impact of words and sentences on the text structure. Word-level and sentence-level of attention mechanisms are adopted to enable differential attention to the more important content when generating the document representation.

**NeurReview: A Neural Architecture Based Conformity Prediction of Peer Review**

Jie Meng

IEEE Access

Year: 2023 | Volume: 11 | Journal Article | Publisher: IEEE

### Peer Review Analysis

As an important paper evaluation mechanism, peer review has been widely adopted in various journals and conferences . Most works on peer review before 2017 were limited to a handful of papers due to the absence of a public domain peer review dataset with sufficient data points. Researchers have explored the usefulness of peer reviews in several aspects based on private review datasets. Xiong et al. examined whether standard product review analysis techniques also apply to our new context of

peer reviews . They also proposed an evaluation system that generates assessments on reviewers' reviewing skills regarding the issue of problem localization. Gender bias in peer-review data has been studied in .

More recently, Kang et al. have collected and analyzed openly available peer review data PeerRead for the first time. They also provided several baselines defining major tasks. Based on this dataset, Wang and Wan have employed peer review text to predict the overall decision status for sentiment analysis and recommendation score prediction by using a Multiple Instance Learning Framework with attention mechanism . Gao et al. [ focused on the role of the rebuttal phase, and proposed a novel task to predict after-rebuttal scores from initial reviews and author responses.

**The Quality Assist: A Technology-Assisted Peer Review Based on Citation Functions to Predict the Paper Quality**

Setio Basuki;Masatoshi Tsuchiya

IEEE Access

Year: 2022 | Volume: 10 | Journal Article | Publisher: IEEE

### Limitattion of Existing Prediction Methods

The literature review poses some limitations in most existing publications. First, the crucial role of *citation functions* was omitted from being addressed in assessing the paper's quality. Second, existing studies did not provide what the manuscript's aspects or sections are important to predict its quality. Third, the unfairness of using review comments as prediction features and using only accuracy as the only metric biased toward the majority class. Fourth, the bias of predicting only *accepted-rejected* due to the final review decision relies on multiple factors. Therefore, this paper develops a prediction method that depends only on the manuscript's content, particularly using the *citation functions* obtained from *citing sentences* to resolve these challenges. We propose creating two additional prediction features, *regular sentences* and *reference-based* features. The paper majorly aims to predict the paper quality (*good-poor*) and the review scores. The final review decision is covered as well for comparison purposes. Accordingly, we address the limitation of determining the most influential part of the manuscript to predict its quality using several ML and FS methods.

Interestingly, the study by conducted experiments on the three classes of accepted, borderline, and rejected, and the two classes accepted and rejected by eliminating the borderline papers. Although eliminating the borderline papers improved the prediction performance, this becomes inapplicable in the entire peer-review process. Additionally, when a reviewer judges a paper as borderline, it does not mean that the other two reviewers judge it as the same since the submitted manuscripts are reviewed by three reviewers and have three different review scores. Due to this reason, we prefer to use the average review scores to determine whether a paper is good or poor (further explanation of this issue is presented in the subsequent section). Casey et al.  proposed good, average, and poor as final quality decisions in which the labels are determined by the annotator and not by conference

**Rating Prediction Model Based on Causal Inference Debiasing Method in Recommendation**

Jiangang Nan;Yajun Wang;Chengcheng Wang

Chinese Journal of Electronics

Year: 2023 | Volume: 32, Issue: 4 | Journal Article | Publisher: CIE

**Recommendation Model Based on Causal Inference**

Although the method of supervised learning has achieved good results, it requires a large amount of training data to cover a variety of recommendations. In practice, online data retention records often follow only one or more recommended policies. It doesn't cover all the recommendations. Therefore, if trained on such samples, the recommended strategies obtained from supervised learning will have a certain bias. Algorithms are more easily influenced a priori by historical policies. In the field of recommendation systems, traditional machine learning can only find correlations between data based on correlations. But after learning the correlation, it cannot give an accurate recommendation result. The purpose of causal inference is to enhance the model's competence to pursue causal effects: it can get rid of the spurious bias, disentangle the desired model effects, and modularize reusable features that generalize well. Although Rubin's framework of potential outcome  is essentially equivalent to Pearl's structural causal model, we use Pearl's structural causal model. Because Pearl's causality can be clearly modeled in our rating prediction model-each node in the structural causal model can be located in the rating prediction model. However, when we cannot model causalities explicitly, we can try Rubin's theory, such as using the propensity ratings . In recent years, in the field of trajectory prediction and computer visio, many researchers have tried to combine causal inference method with

deep neural network to improve model performance. In the recommendation area, there is some work to model recommendation models through causal inference. The clear causalities can improve model transparency. DICE assign users and items with separate embeddings for interest and conformity, and make each embedding capture only one cause by training with cause-specific data which is obtained according to the colliding effect of causal inference. PDA removes the confounding popularity bias in model training and adjusts the recommendation score with desired popularity bias via causal intervention.

**Narrow Convolutional Neural Network for Arabic Dialects Polarity Classification**

Muath Alali;Nurfadhlina Mohd Sharef;Masrah Azrifah Azmi Murad;Hazlina Hamdan;Nor Azura Husin

IEEE Access

Year: 2019 | Volume: 7 | Journal Article | Publisher: IEEE

Proposed the first narrow stacked CNN model called NCNN to classify Arabic tweets into 2, 3, and 5 scales. NCNN consisted model consisted of multiple convolutional and pooling layers trained on top of word embedding. We studied the impact of pooling size, filter size and number of filters on the performance. The proposed model outperforms the benchmark approaches prevalent for Arabic sentiment analysis tasks for a five scale and two and three [ scale. The Arabic Twitter dataset contained a multi-dialect that required exploitation of another structures of the CNN to preserve most significant information from the tweet. The main contributions of this paper were as follows:

1. A multi-layers CNN model was developed to extract and detect a comprehensive representation in Arabic dialect tweets. This was the first model that adopted stacked CNN layers and has been considered as the starting point for various deep neural networks for Arabic dialect twitter sentiment analysis classification. The most existing methods based on hand crafted features are considered as time consuming and laborious. On other hand, deep neural network for Arabic sentiment analysis have been shown to lack deep learning approaches, especially for the Arabic dialect. In this paper, multi, narrow and stacked layers have been used interchangeably to discuss about the proposed model.

2. Sensitivity analysis has been conducted to include the investigations of the effect of the number, the size of the filters and pooling size on the performances of the proposed model.

3. Experiment results have shown that the proposed NCNN model can achieve lower Macro average mean absolute error ($MAE^M$) and higher Macro average recall (P) compared with benchmark approaches.

**Sentiment Classification Based on Part-of-Speech and Self-Attention Mechanism**

Kefei Cheng;Yanan Yue;Zhiwen Song

IEEE Access

Year: 2020 | Volume: 8 | Journal Article | Publisher: IEEE

Sentiment classification is one of the main research topics in Natural Language Processing(NLP), it can be treated as traditional text classification, and solved by some general classification models. Traditional feature engineering-based models usually focus on extracting efficient features such as lexical features, topic-based features. With the rapid development of deep learning, CNN and RNNs have been applied to obtain better representations of sentences for sentiment classification. Especially, Long Short-term Memory Networks (LSTMs) have shown a striking promise in sentiment analysis . Cai and Xia used two individual CNN architectures to learn textual features and visual features, which can be combined as the input of another CNN architecture for exploiting the internal relation between text and image, so as to realize multimedia sentiment analysis. Dong *et al.* put forward an adaptive Recursive Neural Network by modeling syntactic relations on tweet data. Lai *et al.* proposed a Recurrent Convolution Neural Network (RCNN) which uses recurrent structures in the convolution layer to classify texts. Wang *et al.* [ proposed a regional CNN-LSTM model for multi-dimensional sentiment analysis.

More recently, a new research direction in deep learning has emerged, which introduces an attention mechanism to neural network models. The attention mechanism is capable of focusing on the parts of text that are more important to the current task. Hence, various attention-based approaches have been proposed to solve the sentiment analysis task . Wang *et al.*proposed an Attention-based Long Short-Term Memory Network for aspect-level sentiment classification. Hu *et al.* proposed constrained attention networks (CAN) for multi-aspect sentiment analysis, and introduce orthogonal and sparse regularizations to constrain the attention weight allocation, helping learn better aspect-specific sentence representations.

**Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review**

Siti Dianah Abdul Bujang;Ali Selamat;Ondrej Krejcar;Farhan Mohamed;Lim Kok Cheng;Po Chan Chiu;Hamido Fujita

IEEE Access

Year: 2023 | Volume: 11 | Journal Article | Publisher: IEEE

Data-driven in education is a new trend accelerated by global changes lately. The knowledge and insightful information gained from this area provide many advantages that can improve HEI decision-making. To achieve this, educational datasets are collected from various online databases and platforms such as Course Management and Learning Management Systems (LMS) or known as Moodle, Massive Open Online Courses (MOOC), Open Course Ware (OCW), Open Educational Resources (OER), and social media sites such as Twitter, Facebook, YouTube and Personal Learning Environments (PLE) .

Student grade prediction uses machine learning to predict the final score to improve student academic performance by the end of the semester. The aim is to help educators determine the potential students at risk of low results and help them overcome their learning difficulties. Hence, identifying the relevant factors, including student background, academic information, environmental factors, test scores, and Grade Point Average (GPA) or Cumulative Grade Point Average (CGPA), are significant in predicting student performance . However, when a tremendous amount of data is collected and analyzed without being classified in a balanced way, it becomes a significant problem for predicting students' grades.

During the training phases of student grade prediction, imbalanced classification appears when there is an unequal distribution of instances within the target class in the training dataset. Most datasets involve binary classification consisting of two target outputs: the "pass" class as the majority and the "fail" class as a minority. In contrast, some of it comprises more than two different classes, known as multi-class classification. When one class significantly outnumbers the class of the other, the training model usually spends more time processing on the majority classes than the minority ones, which could be less informative. Consequently, it usually leads classifiers to become biased and produce high erroneous. Due to this, many empirical studies are interested in exploring various methods to enhance student grade prediction performance. However, the methods and algorithms used in dealing

with various class-imbalanced distributions to predict student grades are not being highlighted and are not comprehensive enough.

Behavior Prediction of Traffic Actors for Intelligent Vehicle Using Artificial Intelligence Techniques: A Review

Suresh Kolekar;Shilpa Gite;Biswajeet Pradhan;Ketan Kotecha

IEEE Access

Year: 2021 | Volume: 9 | Journal Article | Publisher: IEEE

**Prior Research**

Specifically, in the field of behavior prediction of traffic actors for an intelligent vehicle, as per our knowledge, there are very few Systematic Literature Reviews (SLRs) papers are available. One of the most recent review papers on vehicle behavior prediction for intelligent driving using a deep learning approach was Mozaffari *et al*. In their work, the authors discussed challenges and problems associated with predicting future vehicle trajectories during complex driving scenarios. They provided a comprehensive review of the different approaches used to solve vehicle behavior prediction, i.e., physics-based, maneuver-based, and interaction-aware models. Based on input representation, output type, and prediction model, various researchers have used different approaches. In our view, this work gives a valuable start to researchers who might be interested in vehicle behavior prediction for the safe navigation of intelligent vehicles.

Ridel *et al.* conducted a review in 2018 to predict the behavior of pedestrians in urban scenarios for intelligent vehicles. In this work, the authors discussed the state-of-the-art research developments and challenges to overcome towards finding solutions closer to the human ability to predict and interpret the behavior of pedestrians. This task requires high response time, accuracy, and precision in the real world. However, a lot more research still needs to be done to develop an intelligent vehicle that can ensure the safety of pedestrians on the road.

In a very recent work in 2020, Dunne *et al.* conducted SLR to present the computational model for predicting human behavior in an intelligent environment.

**Optimization and Decomposition Methods in Network Traffic Prediction Model: A Review and Discussion**

Jinmei Shi;Yu-Beng Leau;Kun Li;Yong-Jin Park;Zhiwei Yan

IEEE Access

Year: 2020 | Volume: 8 | Journal Article | Publisher: IEEE

Genetic Algorithm (GA) is a search heuristic algorithm combining genetic and evolutionary computing. It was developed by an American professor, J. Holland, in 1975. In 1989, Goldberg's work made a comprehensive and systematic summary and discussion of the genetic algorithm which laid the foundation for the genetic algorithm which based on the nature of the pattern of change . The purpose of GA is to compare many different individual solutions taken from a population and select the one that fits the data best in terms of a characteristic value. Selection simulates the natural 'law of survival of the fittest.

For the event and the data set contained in the event, Genetic algorithm first collects the characteristic quantity of the data model, and then finds the best individual. The process of genetic algorithm is as follows:

1. The data of the initial population is coded to determine its code length, and the initial population is determined by the random number group output.

2. Calculate the fitness functions of the initial population to determine whether the current requirements are satisfied. If they are, proceed to step Otherwise, go to step .

3. The population in the dataset is selected for replication, crossover and variation, and then output the next generation population. At this time, the number of iterations increases by one, and the genetic algebra increases by one. Then the fitness function of the new population is calculated to determine whether the current requirements are met. If they are, go to step. Otherwise, proceed to step .

4. Return the current optimal individual and complete the entire process.

# CHAPTER 7

# SOURCE CODE

```
import os
import pickle
import streamlit as st
from streamlit_option_menu import option_menu

# Set page configuration
st.set_page_config(page_title="Health Assistant",
                   layout="wide",
                   page_icon=" 🧑‍⚕️ ")



# getting the working directory of the main.py
working_dir = os.path.dirname(os.path.abspath(__file__))

# loading the saved models

diabetes_model = pickle.load(open(f'{working_dir}/saved_models/diabetes_model.sav', 'rb'))

heart_disease_model = pickle.load(open(f'{working_dir}/saved_models/heart_disease_model.sav',
'rb'))

parkinsons_model = pickle.load(open(f'{working_dir}/saved_models/parkinsons_model.sav', 'rb'))

# sidebar for navigation
with st.sidebar:
    selected = option_menu('Multiple Disease Prediction System',
```

```
                ['Diabetes Prediction',
                 'Heart Disease Prediction',
                 'Parkinsons Prediction'],
                menu_icon='hospital-fill',
                icons=['activity', 'heart', 'person'],
                default_index=0)


# Diabetes Prediction Page
if selected == 'Diabetes Prediction':

    # page title
    st.title('Diabetes Prediction using ML')

    # getting the input data from the user
    col1, col2, col3 = st.columns(3)

    with col1:
        Pregnancies = st.text_input('Number of Pregnancies')

    with col2:
        Glucose = st.text_input('Glucose Level')

    with col3:
        BloodPressure = st.text_input('Blood Pressure value')

    with col1:
        SkinThickness = st.text_input('Skin Thickness value')

    with col2:
        Insulin = st.text_input('Insulin Level')
```

```
    with col3:
        BMI = st.text_input('BMI value')

    with col1:
        DiabetesPedigreeFunction = st.text_input('Diabetes Pedigree Function value')

    with col2:
        Age = st.text_input('Age of the Person')



    # code for Prediction
    diab_diagnosis = ''

    # creating a button for Prediction

    if st.button('Diabetes Test Result'):

        user_input = [Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin,
                      BMI, DiabetesPedigreeFunction, Age]

        user_input = [float(x) for x in user_input]

        diab_prediction = diabetes_model.predict([user_input])

        if diab_prediction[0] == 1:
            diab_diagnosis = 'The person is diabetic'
        else:
            diab_diagnosis = 'The person is not diabetic'

    st.success(diab_diagnosis)

# Heart Disease Prediction Page
```

```
if selected == 'Heart Disease Prediction':

    # page title
    st.title('Heart Disease Prediction using ML')

    col1, col2, col3 = st.columns(3)

    with col1:
        age = st.text_input('Age')

    with col2:
        sex = st.text_input('Sex')

    with col3:
        cp = st.text_input('Chest Pain types')

    with col1:
        trestbps = st.text_input('Resting Blood Pressure')

    with col2:
        chol = st.text_input('Serum Cholestoral in mg/dl')

    with col3:
        fbs = st.text_input('Fasting Blood Sugar > 120 mg/dl')

    with col1:
        restecg = st.text_input('Resting Electrocardiographic results')

    with col2:
        thalach = st.text_input('Maximum Heart Rate achieved')

    with col3:
```

```
    exang = st.text_input('Exercise Induced Angina')

with col1:
    oldpeak = st.text_input('ST depression induced by exercise')

with col2:
    slope = st.text_input('Slope of the peak exercise ST segment')

with col3:
    ca = st.text_input('Major vessels colored by flourosopy')

with col1:
    thal = st.text_input('thal: 0 = normal; 1 = fixed defect; 2 = reversable defect')

# code for Prediction
heart_diagnosis = ''

# creating a button for Prediction

if st.button('Heart Disease Test Result'):

    user_input = [age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal]

    user_input = [float(x) for x in user_input]

    heart_prediction = heart_disease_model.predict([user_input])

    if heart_prediction[0] == 1:
        heart_diagnosis = 'The person is having heart disease'
    else:
        heart_diagnosis = 'The person does not have any heart disease'
```

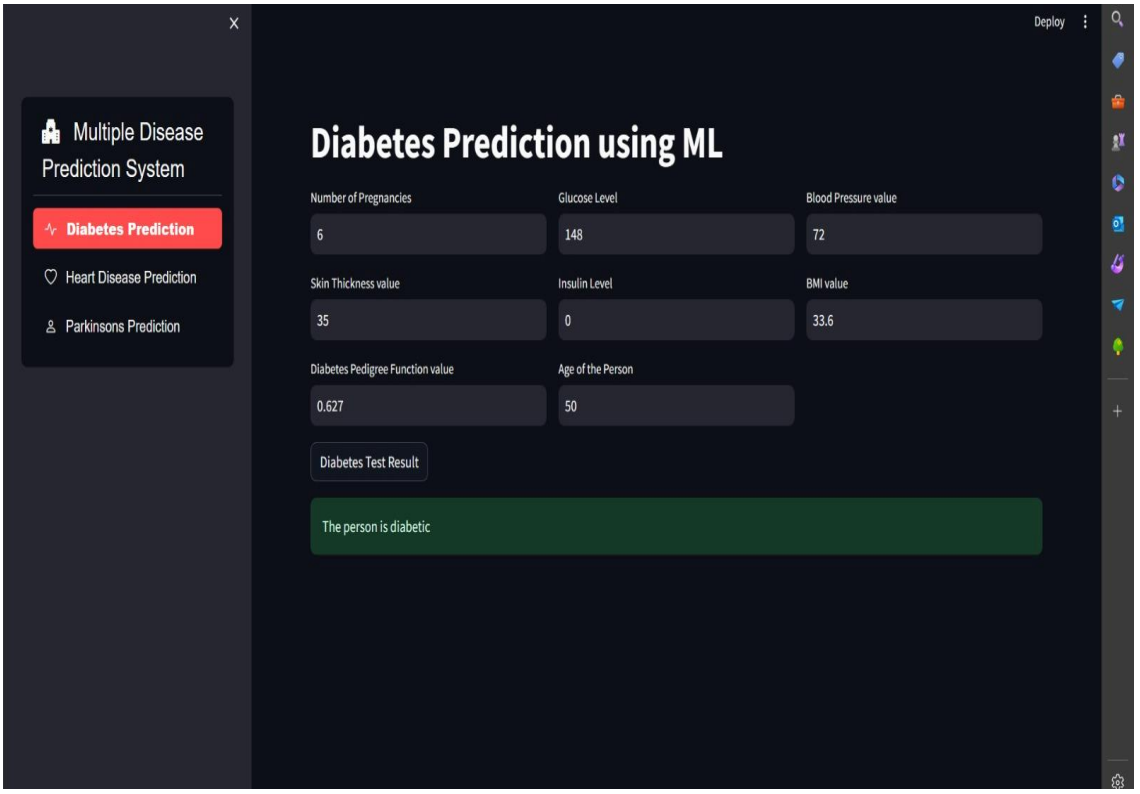
```
    st.success(heart_diagnosis)

# Parkinson's Prediction Page
if selected == "Parkinsons Prediction":

    # page title
    st.title("Parkinson's Disease Prediction using ML")

    col1, col2, col3, col4, col5 = st.columns(5)

    with col1:
        fo = st.text_input('MDVP:Fo(Hz)')

    with col2:
        fhi = st.text_input('MDVP:Fhi(Hz)')

    with col3:
        flo = st.text_input('MDVP:Flo(Hz)')

    with col4:
        Jitter_percent = st.text_input('MDVP:Jitter(%)')

    with col5:
        Jitter_Abs = st.text_input('MDVP:Jitter(Abs)')

    with col1:
        RAP = st.text_input('MDVP:RAP')

    with col2:
        PPQ = st.text_input('MDVP:PPQ')

    with col3:
```

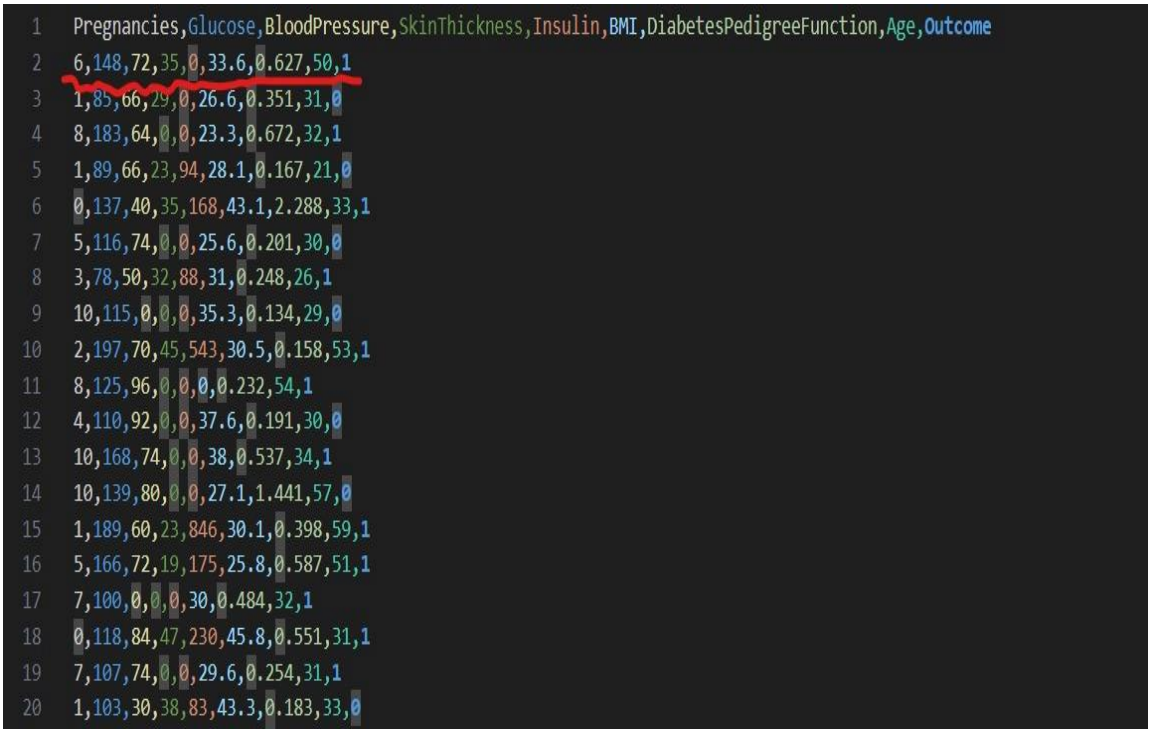
```
    DDP = st.text_input('Jitter:DDP')

with col4:
    Shimmer = st.text_input('MDVP:Shimmer')

with col5:
    Shimmer_dB = st.text_input('MDVP:Shimmer(dB)')

with col1:
    APQ3 = st.text_input('Shimmer:APQ3')

with col2:
    APQ5 = st.text_input('Shimmer:APQ5')

with col3:
    APQ = st.text_input('MDVP:APQ')

with col4:
    DDA = st.text_input('Shimmer:DDA')

with col5:
    NHR = st.text_input('NHR')

with col1:
    HNR = st.text_input('HNR')

with col2:
    RPDE = st.text_input('RPDE')

with col3:
    DFA = st.text_input('DFA')
```

```
with col4:
    spread1 = st.text_input('spread1')

with col5:
    spread2 = st.text_input('spread2')

with col1:
    D2 = st.text_input('D2')

with col2:
    PPE = st.text_input('PPE')

# code for Prediction
parkinsons_diagnosis = ''

# creating a button for Prediction
if st.button("Parkinson's Test Result"):

    user_input = [fo, fhi, flo, Jitter_percent, Jitter_Abs,
                  RAP, PPQ, DDP,Shimmer, Shimmer_dB, APQ3, APQ5,
                  APQ, DDA, NHR, HNR, RPDE, DFA, spread1, spread2, D2, PPE]

    user_input = [float(x) for x in user_input]

    parkinsons_prediction = parkinsons_model.predict([user_input])

    if parkinsons_prediction[0] == 1:
        parkinsons_diagnosis = "The person has Parkinson's disease"
    else:
        parkinsons_diagnosis = "The person does not have Parkinson's disease"

st.success(parkinsons_diagnosis)
```

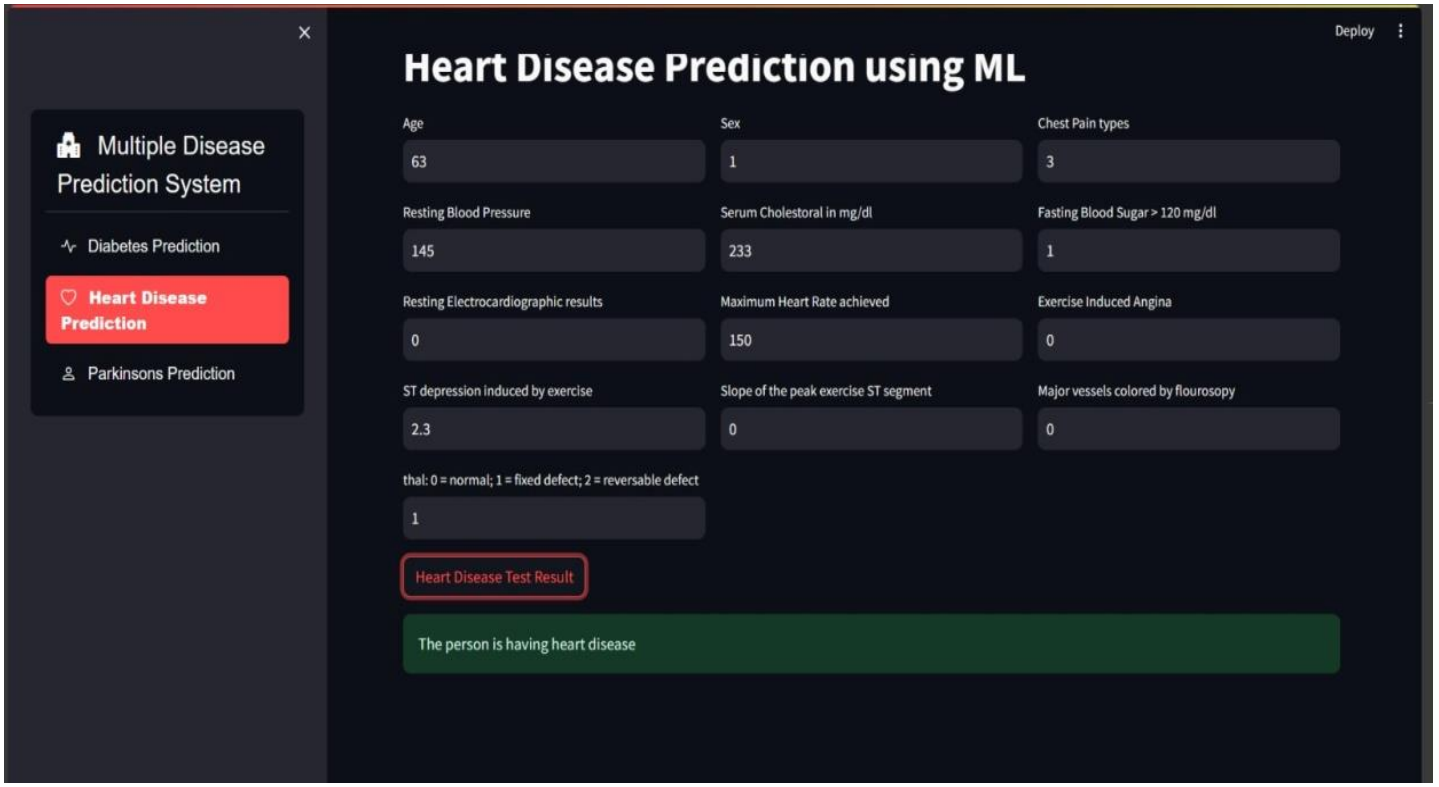
# CHAPTER 8

## Output (Screenshots)



```
Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age,Outcome
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
5,116,74,0,0,25.6,0.201,30,0
3,78,50,32,88,31,0.248,26,1
10,115,0,0,0,35.3,0.134,29,0
2,197,70,45,543,30.5,0.158,53,1
8,125,96,0,0,0,0.232,54,1
4,110,92,0,0,37.6,0.191,30,0
10,168,74,0,0,38,0.537,34,1
10,139,80,0,0,27.1,1.441,57,0
1,189,60,23,846,30.1,0.398,59,1
5,166,72,19,175,25.8,0.587,51,1
7,100,0,0,0,30,0.484,32,1
0,118,84,47,230,45.8,0.551,31,1
7,107,74,0,0,29.6,0.254,31,1
1,103,30,38,83,43.3,0.183,33,0
```




dataset > heart.csv > data

```
age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target
63,1,3,145,233,1,0,150,0,2.3,0,0,1,1
37,1,2,130,250,0,1,187,0,3.5,0,0,2,1
41,0,1,130,204,0,0,172,0,1.4,2,0,2,1
56,1,1,120,236,0,1,178,0,0.8,2,0,2,1
57,0,0,120,354,0,1,163,1,0.6,2,0,2,1
57,1,0,140,192,0,1,148,0,0.4,1,0,1,1
56,0,1,140,294,0,0,153,0,1.3,1,0,2,1
44,1,1,120,263,0,1,173,0,0,2,0,3,1
52,1,2,172,199,1,1,162,0,0.5,2,0,3,1
57,1,2,150,168,0,1,174,0,1.6,2,0,2,1
54,1,0,140,239,0,1,160,0,1.2,2,0,2,1
48,0,2,130,275,0,1,139,0,0.2,2,0,2,1
49,1,1,130,266,0,1,171,0,0.6,2,0,2,1
64,1,3,110,211,0,0,144,1,1.8,1,0,2,1
58,0,3,150,283,1,0,162,0,1,2,0,2,1
50,0,2,120,219,0,1,158,0,1.6,1,0,2,1
58,0,2,120,340,0,1,172,0,0,2,0,2,1
66,0,3,150,226,0,1,114,0,2.6,0,0,2,1
43,1,0,150,247,0,1,171,0,1.5,2,0,2,1
```

# CHAPTER 9

# Conclusion

The potential of Restricted Boltzmann Machines (RBMs) for developing a robust predictive model capable of early disease detection. By leveraging a comprehensive dataset encompassing symptoms, medical history, lifestyle factors, and potentially genetic information, the RBMs were able to extract intricate patterns that hold promise for identifying individuals at risk for various diseases.

This project's significance lies in its potential to empower proactive healthcare interventions. Early disease detection, facilitated by the RBM model, allows for timely interventions that can potentially improve treatment efficacy and mitigate disease severity. By pinpointing significant risk factors and early indicators, the model can guide healthcare professionals towards preventive measures.

The integration of this RBM-based model into clinical decision support systems presents a compelling opportunity to enhance diagnostic accuracy and inform treatment strategies. By providing data-driven insights into disease risk, the system can equip physicians to make more informed decisions, ultimately leading to improved patient outcomes.

It is important to acknowledge that this project serves as a foundation for further exploration. While the results are promising, limitations such as data availability and model complexity necessitate further research. Future endeavors may involve incorporating additional data sources, refining the RBM architecture, and exploring ensemble learning techniques to potentially enhance model performance.

In conclusion, this project has demonstrated the potential of RBMs for early disease prediction. By fostering proactive healthcare strategies and empowering data-driven clinical decision making, this approach holds significant promise for improving patient outcomes and advancing the field of healthcare..

## 9.1 FUTURE WORK

The project's future work encompasses expanding the model to predict more diseases, potentially using Deep Boltzmann Machines. Integration of data from wearables and environmental factors, along with addressing data scarcity and uncertainties, are further areas of exploration. Developing a user interface, exploring clinical integration, and incorporating Explainable AI and personalized medicine approaches hold promise for real-world impact.

The project's future endeavors aim to push the boundaries of disease prediction using RBMs. One key area of exploration involves expanding the model's capabilities to encompass a wider range of diseases. This necessitates not only incorporating additional disease profiles but potentially also exploring the use of Deep Boltzmann Machines (DBMs). DBMs, with their ability to capture more intricate relationships within complex data, could potentially enhance the model's accuracy in disease prediction.

# REFERENCES


[1] A. L. Bui, T. B. Horwich, and G. C. Fonarow, ‘‘Epidemiology and risk profile of heart failure,’’ Nature Rev. Cardiol., vol. 8, no. 1, p. 30, 2011.

[2] M. Durairaj and N. Ramasamy, ‘‘A comparison of the perceptiveapproaches for preprocessing the data set for predicting fertility success rate,’’ Int. J. Control Theory Appl., vol. 9, no. 27, pp. 255–260, 2016.

[3] L. A. Allen, L. W. Stevenson, K. L. Grady, N. E. Goldstein, D. D. Matlock, R. M. Arnold, N. R. Cook, G. M. Felker, G. S. Francis, P. J. Hauptman, E. P. Havranek, H. M. Krumholz, D. Mancini, B. Riegel, and J. A. Spertus,‘‘Decision making in advanced heart failure: A scientific statement from the American heart association,’’ Circulation, vol. 125, no. 15, pp. 1928–1952, 2012.

[4] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, ‘‘Innovative artificial neural networks-based decision support system for heart diseases diagnosis,’’ J. Intell. Learn. Syst. Appl., vol. 5, no. 3, 2013, Art. no. 35396.

[5] Q. K. Al-Shayea, ‘‘Artificial neural networks in medical diagnosis,’’ Int. J.Comput. Sci. Issues, vol. 8, no. 2, pp. 150–154, 2011.

[6] J. Lopez-Sendon, ‘‘The heart failure epidemic,’’ Medicographia, vol. 33, no. 4, pp. 363–369, 2011.

[7] P. A. Heidenreich, J. G. Trogdon, O. A. Khavjou, J. Butler, K. Dracup, M. D. Ezekowitz, E. A. Finkelstein, Y. Hong, S. C. Johnston, A. Khera, D. M. Lloyd-Jones, S. A. Nelson, G. Nichol, D. Orenstein, P. W. F. Wilson, and Y. J. Woo, ‘‘Forecasting the future of cardiovascular disease in the united states: A policy statement from the American heart association,’’ Circulation, vol. 123, no. 8, pp. 933–944, 2011.

[8] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, ‘‘Nonlinearspeech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson’s disease symptom severity,’’ J. Roy. Soc. Interface, vol. 8, no. 59, pp. 842–855, 2011. [14] Langcai Cao;Biyang Ma;Ya Zhou;Bilian Chen,”Design and Implementation of Writing Recommendation system Based on Hybrid Recommendation,”IEEE Access, Vol.no.6,PP. 72506 - 72513,November 2018

[9] D. Liu, Y. Yan, M.-L. Shyu, G. Zhao, and M. Chen, ‘‘Spatio-temporal analysis for human action detection and recognition in uncontrolled environments,’’ Int. J. Multimedia Data Eng. Manage., vol. 6, pp. 1–18, Jan. 2015.

[10] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, ‘‘Effective codebooks for human action representation and classification in unconstrained videos,’’ IEEE Trans. Multimedia, vol. 14, no. 4, pp. 1234–1245, Mar. 2012.

[11] N. Dalal and B. Triggs, ‘‘Histograms of oriented gradients for human detection,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Aug. 2005, pp. 1–9.

[12] R. Cutler and M. Turk, ‘‘View-based interpretation of real-time optical flow for gesture recognition,’’ in Proc. 3rd IEEE Int. Conf. Autom. Face Gesture Recognit., Apr. 1998, pp. 416–421.

[13] D. G. Lowe, ‘‘Distinctive image features from scale-invariant keypoints,’’ Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, 2004.

[14] I. Laptev, ‘‘On space-time interest points,’’ Int. J. Comput. Vis., vol. 64, nos. 2–3, pp. 107–123, 2005.

[15] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, ‘‘An efficient K-means clustering algorithm: Analysis and implementation,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881–892, Jul. 2002.

[16] L. Fei-Fei and P. Perona, ‘‘A Bayesian hierarchical model for learning natural scene categories,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2005, pp. 524–531.

[17] D. Oneata, J. Verbeek, and C. Schmid, ‘‘Action and event recognition with Fisher vectors on a compact feature set,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2013, pp. 1817–1824.

[18] M. Bregonzio, S. Gong, and T. Xiang, ‘‘Recognising action as clouds of space-time interest points,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 1948–1955.

[19] M. S. Ryoo and J. K. Aggarwal, ‘‘Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Sep. 2009, pp. 1–2.

[20] A. Kovashka and K. Grauman, ‘‘Learning a hierarchy of discriminative space-time neighborhood features for human action recognition,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2046–2053.

[21] J. Wen, Z. Lai, Y. Zhan, and J. Cui, ‘‘The L2,1-norm-based unsupervised optimal feature selection with applications to action recognition,’’ Pattern Recognit., vol. 60, pp. 515–530, Dec. 2016. [22] G. Varol and A. A. Salah, ‘‘Efficient large-scale action recognition in videos using extreme learning machines,’’ Expert Syst. Appl., vol. 42, no. 21, pp. 8274–8282, 2015.

[23] Y. Yuan, L. Qi, and X. Lu, ‘‘Action recognition by joint learning,’’ Image Vis. Comput., vol. 55, pp. 77–85, Nov. 2016.

[24] W. Xu, Z. Miao, and X.-P. Zhang, ‘‘Structured feature-graph model for human activity recognition,’’ in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2015, pp. 1245–1249.

[25] Y. Wang and G. Mori, ‘‘Max-margin hidden conditional random fields for human action recognition,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 872–879.

[26] Y. Wang and G. Mori, ‘‘Learning a discriminative hidden part model for human action recognition,’’ in Proc. Adv. Neural Inf. Process. Syst., 2009, pp. 1721–1728.

[27] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell, ‘‘Hidden conditional random fields,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 10, pp. 1848–1852, Aug. 2007.

[28] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, ‘‘Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,’’ in Proc. 18th Int. Conf. Mach. Learn., 2001, pp. 1225–1265.

[29] Y. M. Liang, S. W. Shih, A. C. C. Shih, H. Y. M. Liao, and C. C. Lin, ‘‘Learning atomic human actions using variable-length Markov models,’’ IEEE Trans. Syst., Man, B (Cybern.), vol. 39, no. 1, pp. 268–280, Feb. 2009.

[30] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, “Prediction and validation of disease genes using HeteSim scores,” IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 14, no. 3, pp. 687–695, May–Jun. 2017.
[31] S. B. U. Martin, N. Nagarajan, T. Ambuj, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, “Correction: Prediction and validation of gene–disease associations using methods inspired by social network analyses,” PLoS One, vol. 8, no. 5, p. e58977, 2013.

[32] S. Ballouz et al., “Candidate disease gene prediction using Gentrepid: Application to a genome-wide association study on coronary artery disease,” Mol. Genetics Genomic Med., vol. 2, no. 1, pp. 44–57, 2014.

[33] S. Elshal, L. C. Tranchevent, A. Sifrim, A. Ardeshirdavani, J. Davis, and Y. Moreau, “Beegle: From literature mining to disease-gene discovery,” Nucleic Acids Res., vol. 44, no. 2, pp. e18–e18, 2016.

[34] D. Gligorijevic et al., “Large-scale discovery of disease–disease and disease–gene associations,” Sci. Rep., vol. 6, 2016, Art. no. 32404.

[35] Y. Luo et al., “A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information,” Nature Commun., vol. 8, no. 1, 2017, Art. no. 573.

[36] J. Pinero ˜ et al., “DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes,” Database: J. Biol. Databases Curation, vol. 2015, Art. no. bav028, 2015.

[37] N. Rappaport et al., “MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search,” Nucleic Acids Res., vol. 45, no. Database issue, pp. D877–D887, 2017.

[38] J. Menche et al., “Disease networks. Uncovering disease–disease relationships through the incomplete interactome,” Science, vol. 347, no. 6224, p. 1257601, 2015.

[39] S. Kohler ¨ et al., “The human phenotype ontology in 2017,” Nucleic Acids Res., vol. 45, no. Database issue, pp. D865–D876, 2017.

[40] A. Rath et al., “Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users,” Human Mutation, vol. 33, no. 5, pp. 803–808, 2012.