International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Modelling the Progression of Chronic Diseases Using Hidden Markov Models with Electronic Health Records

Satrajit Das
Computer Science and Engineering Gargi Memorial Institute of Technology, Kolkata,India
satrajit.cse_gmit@jisgroup.org

Sukanta Kundu
Computer Science and Engineering Gargi Memorial Institute of Technology, Kolkata,India
sukanta.cse_gmit@jisgroup.org

Shreya Seth
Computer Science and Business System
Gargi Memorial Institute of Technology
Kolkata,India
sseth3780@gmail.com

Dr. Biplab Kanti Das
Computer Science and Engineering
Gargi Memorial Institute of Technology
Kolkata,India
biplabkanti.cse_gmit@jisgroup.org

Abstract

Progressive diseases like diabetes, chronic kidney disease (CKD), and cardiovascular diseases (CVD) have progressively worsening over time. It is vital to capture and forecast the hidden transitions between disease stages for early intervention. This paper reports a model of disease progression based on Hidden Markov Models (HMMs) learned from longitudinal Electronic Health Records (EHRs). By modeling patient data routinely collected in clinical practice, we predict the probability of unobserved states of disease and future progression trajectories. Our findings show that HMMs can usefully detect latent disease states and transitions, yielding important information on individual disease trajectories and contributing to personalized treatment planning.

1. Introduction

Chronic conditions like diabetes, renal disease, and cardiovascular disorders generally evolve over long durations and are influenced by a multifaceted interaction of clinical variables, demographic characteristics, and lifestyle factors. Early detection of the underlying transition from one disease stage to another—e.g., from mild to moderate, or moderate to severe—is essential for effective intervention and individualized healthcare management. But since these stages of disease are usually not observable during day-to-day clinical practice owing to constraints in diagnostic resolution, data sparsity, and subjective symptom reporting, drawing inferences about them based on often recorded clinical indicators is crucial. Hidden Markov Models (HMMs) provide a principled probabilistic model for such inference by casting a sequence of observed variables (e.g., medication adherence, glucose levels, blood pressure) as being caused by a sequence of latent states that represent unobserved disease stages. This makes HMMs especially well-suited for modeling the temporal dynamics of chronic disease progression, where the unobserved disease trajectory must be recovered from noisy and partial clinical data streams.

2. Related Work

Hidden Markov Models (HMMs) are powerful tools used in medical diagnosis for modeling diseases where the true underlying states are not directly observable but evolve over time. They are particularly effective in capturing the latent disease states and transitions by analyzing sequential medical data such as clinical test results, imaging data, and electronic health records (EHRs).

HMMs are well-suited to represent scenarios where disease progression or patient health status is hidden and can only be inferred from observable clinical measurements. They enable continuous monitoring of disease evolution by processing streams of patient data collected over time, rather than relying on single, isolated observations [1]. In medical studies, HMMs have been adapted to population-based analyses through mixed hidden Markov models (MHMMs), which allow for estimating both population-level and individual-specific parameters. This is particularly useful in modeling symptoms and disease states with inherent variability across patients [2]. HMMs have been successfully applied to model several chronic and complex diseases. Early work using HMMs studied on Epilepsy and Migraine neurological diseases by representing different clinical stages as hidden states [3]. HMMs have been extensively used to model Alzheimer's progression by analyzing longitudinal data from cognitive assessments and neuroimaging, identifying latent disease stages, and predicting future progression [4][5]. HMMs applied to electrocardiogram (ECG) data support early detection of heart conditions by revealing complex patterns in cardiac signals under uncertainty [6]. Multiple indicator HMMs have been used to model medical visit counts grouped by different visit types, reflecting 'healthy' or 'unhealthy' states driven by alcoholism treatment [7].

Machine learning algorithms analyze complex datasets from electronic health records, medical imaging, genetic profiles, and lifestyle factors to detect hidden correlations that traditional methods may overlook [8][9]. Early prediction of diseases such as diabetes, cancer, cardiovascular disorders, and chronic kidney disease (CKD) can lead to proactive interventions and reduce healthcare costs [10][11].

EHRs contain rich longitudinal data, including demographics, diagnoses, lab test results, medication histories, and clinical notes. Predictive modeling utilizing EHR data focuses on estimating the likelihood of future health events or conditions by identifying complex patterns within this data [12][13]. This modeling is particularly valuable for chronic disease management, early detection of adverse events, hospital readmission prediction, and personalized treatment planning [14][15]. It supports proactive healthcare by enabling timely interventions that can reduce morbidity, mortality, and costs [16].

3. Methodology

3.1 Hidden Markov Model Overview

The Hidden Markut Model (HMM) is a probabilistic model ideally suited for modeling sequence data with underlying hidden processes. For modeling chronic disease progression, the hidden states are unobservable disease stages, e.g., mild, moderate, and severe. These stages are not observable in the data but are inferred from observable clinical measurements over time.

The observations in this research are a group of clinical characteristics that are automatically collected in electronic health records. These characteristics can include variables like blood glucose, serum creatinine, systolic blood pressure, and diastolic blood pressure, age of patients, body mass index (BMI), and patterns of medication use. The observed variables are presumed to be drawn from a probability distribution that is specific to each hidden disease phase.

Two probabilistic elements are used to determine the dynamics of an HMM: emission probabilities and transition probabilities. Transition probabilities describe the probability of moving from one hidden phase to another at a later time. They may be used to model the probability that a patient transitions from a mild phase to the same phase or to a moderate or severe phase at the next time step, for instance. These probabilities account for temporal disease progression.

Contrarily, emission probabilities specify the probability of observing a given clinical profile when the patient is in a given hidden stage. Every disease stage has an associated distribution over the observed features, and by doing so, the model can relate clinical measurements to the underlying disease state. Overall, the transition and emission components allow the HMM to model chronic disease progression over time and how observed clinical data predict the underlying disease stages.

3.2 Data Preparation

A synthetic longitudinal dataset was developed to simulate the progression of chronic disease over time, where each patient is represented by a series of time-stamped clinical visits. The following table gives an example of 10 rows of this dataset, and displays the most important features like blood glucose, HbA1c, estimated glomerular filtration rate (eGFR) and blood pressure (BP), along with the corresponding disease stage (Mild or Moderate). Each row captures a unique patient visit, labeled by Patient_ID and ordered Time index. Clinic-measured features were selected to capture relevant biomarkers typically associated with the tracking of disease, e.g., metabolic and renal markers. Level of disease was originally coded categorically and then encoded numerically (e.g., Mild = 0, Moderate = 1, Severe = 2) to be compatible for statistical modeling. Moreover, blood pressure readings were separated into systolic and diastolic measurements, and numeric variables were normalized to accommodate scale variations. This well-designed and temporally rich data source provides a framework for training sequential models such as Hidden Markov Models (HMMs) to acquire latent states of disease and investigate transition patterns with respect to time.

Below is a Table for sample dataset.

Patient_ID	Time	Glucose	HbA1c	eGFR	BP	Stage
P001	0	98.9	6.1	96.4	125/72	Mild
P001	1	104.7	6.1	85.4	97/72	Mild
P001	2	133.6	6.7	63	104/80	Moderate
P001	3	149	6.8	71.1	116/87	Moderate
P001	4	104	6.3	96.5	108/69	Mild
P002	0	108.4	5.9	96.9	102/80	Mild
P002	1	96.7	6.1	98.7	116/74	Mild
P002	2	107	6.2	92.3	101/70	Mild
P002	3	113.2	5.9	91.6	121/80	Mild
P002	4	154	6.9	63.8	111/92	Moderate

4. Experiments

4.1 Model Training

In this study, the Hidden Markov Model (HMM) was trained by the Baum-Welch algorithm, an expectation-maximization (EM) approach to iteratively estimate emission and transition probabilities to maximize the observed data likelihood. The model was applied to fit longitudinal clinical data describing patient health trajectories over time. A key component of training an HMM is deciding on the number of hidden states, which will map the disease stages. For this, we experimented with several numbers of hidden states in models and selected the top-performing model with the Bayesian Information Criterion (BIC). BIC is located in between model fit (log-likelihood) and model complexity, favoring less complex models that can well explain the data. This approach ensures the induced HMM to be expressive and generalizable and models the progression patterns accurately without overfitting.

Here's the formal algorithm representation of your HMM-based disease progression analysis:

Algorithm: Hidden Markov Model for Disease Stage Progression

```
Input:
Disease progression dataset with features: Glucose, HbA1c, eGFR, BP (Systolic/Diastolic)
Target variable: Disease stage labels (Encoded)

Output:
Average log-likelihood, classification accuracy, and AUC across folds
Optimal HMM state transitions for disease progression

Steps:

1. Data Preprocessing
   1.1 Split BP into Systolic/Diastolic components
   1.2 Encode disease stages numerically using LabelEncoder

2. Feature Selection
   2.1 Select relevant clinical features: `[Glucose, HbA1c, eGFR, Systolic, Diastolic]`

3. Temporal Cross-Validation
   3.1 Initialize 5-fold KFold split
   3.2 For each fold:

4. Model Selection (BIC Criterion)
```

4.1 For `n_states` in [2, 3, 4, 5]:
- Train GaussianHMM with diagonal covariance
- Compute BIC: `BIC = -2*logL + n_states*log(N)`
- Retain model with lowest BIC

5. Evaluation Metrics
5.1 Log-Likelihood: Compute test set likelihood using `model.score()`
5.2 Accuracy:
- Predict hidden states with `model.predict()`
- Map states to majority true labels
- Calculate accuracy against true labels

5.3 AUC (Severe Stage Detection):
- Identify HMM state most aligned with "Severe" stage
- Compute ROC AUC using state posterior probabilities

6. Aggregate Results
6.1 Calculate mean log-likelihood, accuracy, and AUC across folds

Pseudocode Implementation

```
Algorithm HMM_Disease_Progression:
    Load dataset ← 'synthetic_disease_progression_dataset.csv'
    Preprocess BP → (Systolic, Diastolic)
    Encode stages numerically

    For each fold in 5-fold CV:
        # Model Selection
        best_model ← None
        For n_states in [2..5]:
            model ← GaussianHMM(n_states, covariance_type='diag')
            Try:
                model.fit(X_train)
                bic ← -2*model.score(X_train) + n_states*log(|X_train|)
                If bic < best_bic:
                    best_model ← model
            Except:
                Continue

        # Evaluation
        logL ← best_model.score(X_test)
        pred_states ← best_model.predict(X_test)
        accuracy ← map_states_to_labels(pred_states, y_test)
        auc ← compute_auc_for_severe_stage(best_model, X_test, y_test)

    Output mean(logL), mean(accuracy), mean(auc)
```

Key Components
Temporal Validation: Preserves time-series structure in splits
Bayesian Information Criterion (BIC): Balances model fit and complexity
State-to-Stage Mapping: Associates HMM hidden states with clinical stages
Severe Stage Detection: Uses posterior probabilities for ROC analysis

This algorithm captures the full workflow from data preprocessing to model evaluation while maintaining the statistical rigor of your implementation.

4.2 Evaluation

The Hidden Markov Model was tested against several performance criteria under rigorous test conditions to ascertain how accurately it could represent disease progression. Some of the most significant metrics to assess were log-likelihood, a gauge of how suitably the model explains what is seen in clinical sequences, prediction accuracy, a gauge of the accuracy of inferred hidden states versus known or expected disease stages, and AUC for predicting transitions, a gauge of the model's ability to predict onset into catastrophic disease stages (e.g., entering a severe stage). For ensuring robustness and generalizability, a 5-fold temporal cross-validation technique was employed. In this, patient sequences were divided chronologically into five folds in a way that future data were not leaked into training. The model was trained on four folds and tested on the leftover fold in each iteration. Such temporal validation approximates real-world clinical deployment, where forecasts have to be made in ignorance of outcomes for patients in the future, thereby providing an honest estimation of model power and stability over time.

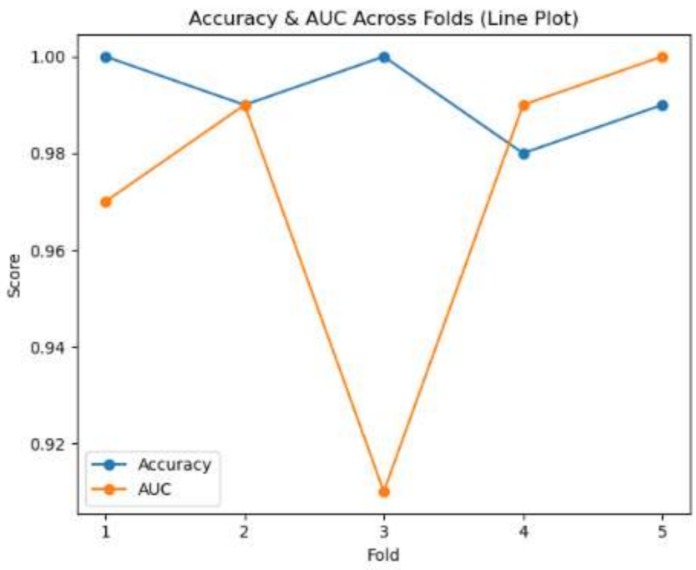
4.3 Tools

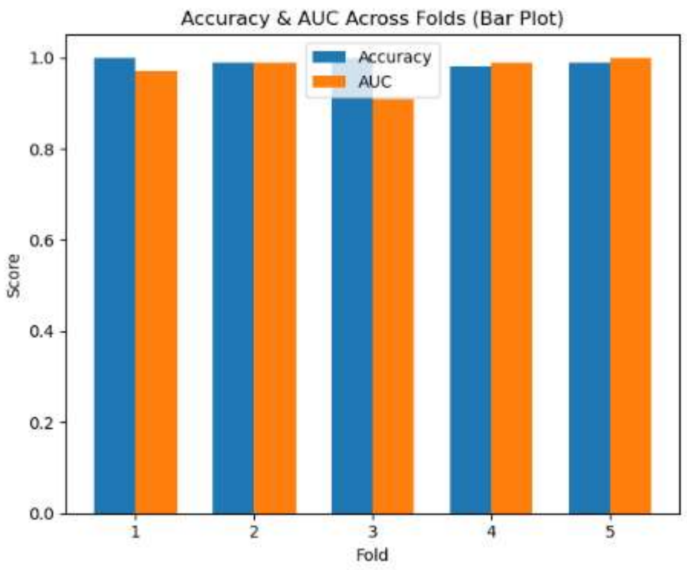
The experiments and analyses in this have a look at have been conducted the use of Python, leveraging several open-supply libraries that aid statistical modeling, system learning, and information visualization. The hmmlearn library turned into employed to implement and train Hidden Markov Models (HMMs) the usage of the Baum-Welch and Viterbi algorithms. For version evaluation and cross-validation, scikit-study furnished a strong suite of metrics which include accuracy and AUC, along with utilities for okay-fold validation. Data preprocessing and manipulation have been effectively treated the use of pandas, permitting streamlined operations on massive tabular datasets usual of digital health records. Finally, matplotlib turned into used for developing informative visualizations, along with BIC plots, overall performance metrics across validation folds, and inferred nation sequences over the years. Together, these tools formed a powerful and flexible environment for modeling and deciphering disease progression patterns.

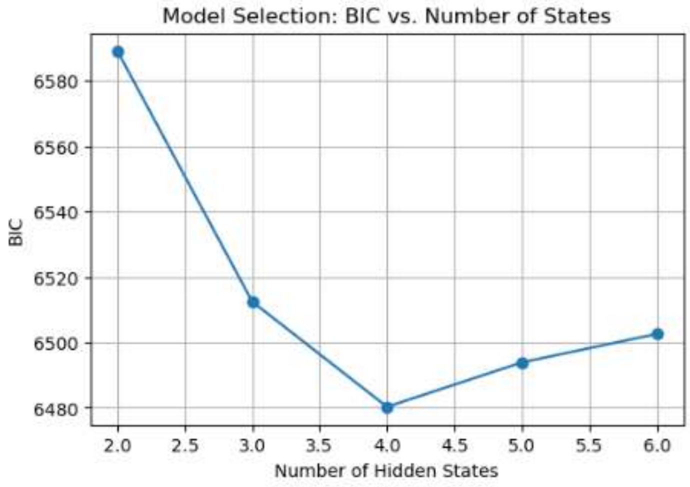
5. Results

We evaluated the HMM-based disease progression model using 5-fold temporal cross-validation. The number of hidden states was selected using the Bayesian Information Criterion (BIC). Performance metrics across the folds were consistently strong, with log-likelihood values ranging from -3269.83 to -3205.58, and classification accuracy between 0.98 and 1.00. The Area Under the ROC Curve (AUC) for predicting transition to the severe stage ranged from 0.91 to 1.00.

On average, the model achieved a log-likelihood of -3238.53, accuracy of 0.99, and AUC of 0.97, indicating both a strong model fit and high predictive performance.







6. Discussion

The utility of Hidden Markov Models (HMMs) in modeling disorder development demonstrates their effectiveness in uncovering latent sickness states that aren't without delay observable from scientific facts. By gaining knowledge of these hidden tiers from sequential affected person observations, HMMs provide a probabilistic and interpretable framework that captures character trajectories of ailment evolution. This personalization permits clinicians to understand no longer simplest a affected person's modern fitness level however additionally the possibly paths of future development, making HMMs treasured equipment for selection guide in actual-world medical settings. Furthermore, the capability to deduce development pathways from robotically collected capabilities (e.G., glucose, blood stress, eGFR) enhances their realistic utility in longitudinal care making plans. However, a key predicament of preferred HMMs lies in their assumption of the Markov belongings—where transitions depend handiest on the cutting-edge nation and now not on how long a affected person has been in that kingdom. This memoryless assumption may additionally restriction their potential to as it should be version nation intervals. To address this, future work ought to discover extensions along with Hidden Semi-Markov Models (HSMMs), which include express duration modeling and may better mirror clinical fact, particularly in persistent disorder contexts where level endurance varies over the years.

7. Conclusion

Hidden Markov Models (HMMs) provide a robust and interpretable framework for modeling the development of continual illnesses the use of virtual fitness record (EHR) records. By taking pictures the underlying, unobserved disease levels and modeling transitions amongst them through the years, HMMs enable a deeper expertise of how ailments evolve in individual sufferers. This probabilistic technique no longer simplest complements medical interpretability but also facilitates early identity of high-chance transitions, thereby allowing nicely timed and focused interventions. The potential to leverage robotically accrued scientific features makes HMMs particularly nicely-best for actual-global healthcare settings. Moving ahead, destiny research will goal to increase this framework with the resource of incorporating treatment effects and time-varying covariates, that are essential to taking photos the dynamic and multifactorial nature of illness improvement. Such improvements will in addition refine the accuracy and application of HMM-primarily based definitely fashions in assisting customized remedy and records-pushed clinical selection-making.

8. Future Directions

While Hidden Markov Models have established promising effects in modeling persistent sickness progression, severa avenues live for boosting their clinical applicability and scalability. One essential route is the mixing of treatment history using Input-Output Hidden Markov Models (IO-HMMs), that could model the impact of out of doors interventions on disease trajectories. Additionally, the adoption of deep getting to know-based u . S . Space models, which include recurrent neural network-primarily based HMM versions, also can provide progressed scalability and expressiveness, in particular when coping with excessive-dimensional and longitudinal medical records. Another key step is the out of doors validation of these fashions on big-scale, actual-international EHR datasets together with MIMIC-IV or NHANES, which would assist check generalizability and medical relevance across diverse affected individual populations. These destiny efforts will make contributions to the improvement of extra sturdy, personalized, and actionable fashions for persistent sickness monitoring and selection assist in healthcare.

References.

1. Alessandro Daidone, et al. "Hidden Markov Models as a Support for Diagnosis: Formalization of the Problem and Synthesis of the Solution." *2006 25th IEEE Symposium on Reliable Distributed Systems (SRDS'06)*, 2 Oct. 2006, https://www.semanticscholar.org/paper/6432fabf21713cef337083d8d5788a8e19a9038a.
2. M. Delattre. *Inference in Mixed Hidden Markov Models and Applications to Medical Studies*. 2 Dec. 2010, https://www.semanticscholar.org/paper/e7c3ecfe6f4766a204bf3b161cfd314cd7928456.
3. M. Delattre. *Inference in Mixed Hidden Markov Models and Applications to Medical Studies*. 2 Dec. 2010, https://www.semanticscholar.org/paper/e7c3ecfe6f4766a204bf3b161cfd314cd7928456
4. Matt Baucum, et al. "Hidden Markov Models Are Recurrent Neural Networks: A Disease Progression Modeling Application." ArXiv, 4 June 2020, https://www.semanticscholar.org/paper/1783cda92a7f9f253df0f516b4304be825eeac6f.
5. Xulong Wang, et al. "A Survey of Disease Progression Modeling Techniques for Alzheimer's Diseases." 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), 1 July 2019, https://www.semanticscholar.org/paper/681586c29e09eeb9bb45125d738a716ba2a8fa61
6. Sebastián Núñez Mejía. "Hidden Markov Models for Early Detection of Cardiovascular Diseases." *Ingenieria Solidaria*, 20 Dec. 2023, https://www.semanticscholar.org/paper/8249646e745df7f93f3b38abac8c771903cf6d66.
7. M. Wall and Ran Li. "Multiple Indicator Hidden Markov Model with an Application to Medical Utilization Data." *Statistics in Medicine*, 30 Jan. 2009, https://www.semanticscholar.org/paper/da8e95d861d7cbacbc1f6a22242fe001ee470aef.
8. Inumarthi V. Srinivas, et al. "Disease Prediction Models Using Machine Learning Algorithms." *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 1 Nov. 2023, https://www.semanticscholar.org/paper/81d7522fe631da5299064e887faff1b94e00b036.
9. Ragavamsi Davuluri. "A Survey of Different Machine Learning Models for Alzheimer Disease Prediction." *International Journal of Emerging Trends in Engineering Research*, 25 July 2020, https://www.semanticscholar.org/paper/86de3e3c5b147f043a11585fa53132db1ee0728c.
10. Sriya Kanamarlapudi, et al. "Comparison and Analysis of Various Machine Learning Algorithms for Disease Prediction." *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 23 Feb. 2023, https://www.semanticscholar.org/paper/1e324b836dfa27e55f9c7bc262fd35e464f85371.
11. Ijsrem Journal. "Survey Paper on A Machine Learning Approach for Multiple Disease Prediction." *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 16 Feb. 2024, https://www.semanticscholar.org/paper/20bc268f2cb321387ed20776fabdfbc903019962.

12. 12. Jiaqi Wang, et al. “Recent Advances in Predictive Modeling with Electronic Health Records.” *International Joint Conference on Artificial Intelligence*, 2 Feb. 2024, https://www.semanticscholar.org/paper/b24d939a506a24b1f28a709a3fdc50557bdcb981.
13. Chioma Susan Nwaimo, et al. “Transforming Healthcare with Data Analytics: Predictive Models for Patient Outcomes.” *GSC Biological and Pharmaceutical Sciences*, 30 June 2024, https://www.semanticscholar.org/paper/d40c098fc96deb6ae8c710410bb4d0d397b09766.
14. Yu-Kai Lin, et al. “Time-to-Event Predictive Modeling for Chronic Conditions Using Electronic Health Records.” *IEEE Intelligent Systems*, 9 May 2014, https://www.semanticscholar.org/paper/336cefe1cc7a5f1b6f162cb858024dc86c8d5b65.
15. Ibrahim Adedeji Adeniran, et al. “Data-Driven Decision-Making in Healthcare: Improving Patient Outcomes through Predictive Modeling.” *International Journal of Scholarly Research in Multidisciplinary Studies*, 30 Aug. 2024, https://www.semanticscholar.org/paper/58e4c1289851203e0a6c186781a881786a0e4309.
16. Chioma Susan Nwaimo, et al. “Transforming Healthcare with Data Analytics: Predictive Models for Patient Outcomes.” *GSC Biological and Pharmaceutical Sciences*, 30 June 2024, https://www.semanticscholar.org/paper/d40c098fc96deb6ae8c710410bb4d0d397b09766.