Medical Diagnosis using Probabilistic AI: Bayesian Networks and Gaussian Processes for Predicting Dementia and Parkinson's Disease

Alfie Atkinson
University of Lincoln
25715017@students.lincoln.ac.uk

***Abstract*—This paper explores the application of probabilistic artificial intelligence models for medical diagnosis, specifically focusing on dementia and Parkinson's disease prediction. We compare the performance of three models: Discrete Bayesian Networks (DBNs), Gaussian Bayesian Networks (GBNs), and Gaussian Processes (GPs), evaluating their ability to handle both discrete and continuous data. Performance metrics such as accuracy, F1-score, AUC, Brier score, and log loss are used to assess model effectiveness. Our findings indicate that DBNs provide superior performance in terms of accuracy and computational efficiency, while GBNs and GPs offer distinct advantages in handling continuous data. This study highlights the strengths, limitations, and computational costs of each model, offering recommendations for their application in medical diagnostics.**

I. Introduction

This report compares probabilistic machine learning models for predicting dementia and Parkinson's disease facilitated by the supporting code [1]. The models evaluated are Discrete Bayesian Networks (DBNs), Gaussian Bayesian Networks (GBNs), and Gaussian Processes (GPs), assessing their predictive performance on real-world datasets with both discrete and continuous features [2], [3].

The report aims to:

- Compare the accuracy of DBNs, GBNs, and GPs.
- Evaluate performance using metrics like accuracy, F1-score, AUC, Brier score, log loss, and time.
- Examine how each model handles discrete and continuous data.
- Discuss strengths, weaknesses, and computational costs.
- Offer recommendations based on findings.

II. Background & Theory

A. Bayesian Networks

Bayesian Networks (BNs) are probabilistic graphical models using directed acyclic graphs (DAGs) to represent conditional dependencies [4], [5]. They are useful for modeling joint probability distributions and making predictions.

Gaussian Bayesian Networks (GBNs) extend BNs to handle continuous variables with Gaussian distributions, suited for applications like medical diagnosis [6], [7].

B. Gaussian Processes

Gaussian Processes (GPs) are non-parametric methods for regression, classification, and optimisation, offering flexibility in modeling non-linear relationships with uncertainty estimates [8], [9]. They face computational challenges but can be addressed with approximation techniques like sparse and Vecchia approximations [10].

C. Comparison of Models

BNs are widely used in healthcare for various medical conditions [11]. While effective for discrete data, they struggle with continuous variables, which GBNs address. GPs extend BNs' capabilities by modeling non-linear dependencies [12].

III. Data Pre-Processing & Dataset Analysis

The steps noted in this section can be found in the Jupyter notebook `dataset-processing-analysis.ipynb` in the GitHub repository for this project [1].

A. Dataset Overview

The datasets used in this study are related to two medical conditions: dementia [2] and Parkinson's disease [3]. Both datasets contain a combination of discrete and continuous variables, which makes them suitable for testing the performance of DBN, GBNs, and GPs.

B. Pre-Processing Steps

Several pre-processing steps were necessary to prepare the data for model training:

- **Target Variable Examination**: Analysed the target variables to identify class imbalances.
- **Data Type Handling**: Categorical variables in the dementia dataset were converted to numerical values.
- **Missing Value Imputation**: K-Nearest Neighbours imputation was applied to the 'SES' variable in the dementia data, where values were missing for 'Demented' patients.
- **Addressing Class Imbalance**: Synthetic Minority Oversampling Technique (SMOTE) was used to balance the class distribution in both datasets.
- **Discretisation for Discrete BNs**: Continuous features in the datasets were discretised using the Freedman-Diaconis rule to prepare for DBNs.

- **Data Shuffling and Saving**: The datasets were shuffled to randomise the order of instances and saved as new CSV files to prevent bias and avoid repeating pre-processing steps.

C. Exploratory Data Analysis

An Exploratory Data Analysis (EDA) was performed to understand the dataset's structure and relationships between variables, employing several visualisation and statistical techniques:

- **Target Variable Examination**: The distribution of the targets was examined. Ambiguous categories like 'Converted' in the dementia dataset were resolved for clarity.
- **Skewness Analysis**: The skewness of features was calculated and compared between classes to identify features that best distinguish between classes.
- **Correlation Analysis**: Correlation matrices were created to find relationships between features and the target. Strong correlations were observed for some features.
- **Histogram Analysis**: Histograms for all features were created, showing the overall distribution of each.

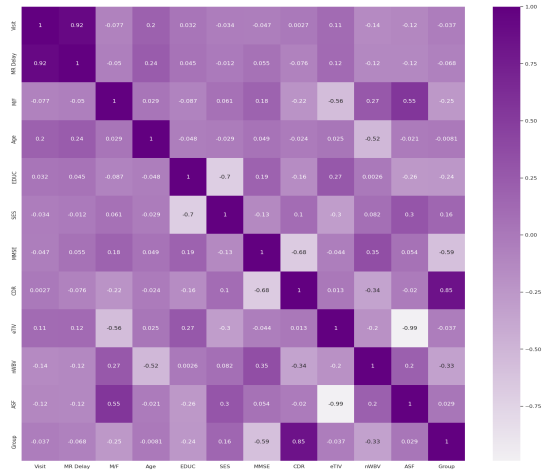

Fig. 1. Correlation matrix of the dementia dataset.

D. Exploratory Data Analysis Results

The correlation analysis revealed that features like cognitive test scores and age were strongly correlated with disease status, making them potentially useful predictors. Additional charts can be found in the appendices, with further insight being available in the relevant notebook [1].

IV. Methodology

A. Library Selection

The following libraries were selected for their robust functionalities and compatibility with the data and models used in this study:

- **`pgmpy`**: Used for structure learning and parameter estimation in DBNs and GBNs.
- **`scikit-learn`**: Used for data pre-processing, model evaluation, and implementing GPs.
- **`matplotlib`** and **`seaborn`**: Used for data visualisation.

B. Model Setup & Training

1) Discrete Bayesian Network (DBN):

- **Data Discretisation**: Sturges' Formula and the Freedman-Diaconis Rule showed the lowest Mean Squared Error for the dementia and Parkinson's dataset, respectively.
- **Structure Learning**: Tested Hill Climbing (BDeu scoring), PC Algorithm, and Tree-Augmented Naive Bayes (TANB) method. Chose TANB for its reliability in including the target variable.
- **Parameter Learning**: Used Bayes estimation for CPDs, showing superior results despite longer computation time compared to Maximum Likelihood Estimation (MLE).

2) Gaussian Bayesian Network (GBN):

- **Structure Learning**: Tested Hill Climbing (BDeu scoring) and PC Algorithm. The PC Algorithm did not reliably include the target.
- **Parameter Learning**: Used MLE to estimate Gaussian distributions.

3) Gaussian Process (GP):

- **Kernel Selection**: Tested Radial Basis Function (RBF) kernel and Matern kernel, with similar results.
- **Parameter Tuning**: Used grid search with cross-validation to optimise kernel parameters.
- **Training**: Trained on the training set and evaluated on the test set.

C. Model Evaluation Criteria

The models were evaluated based on:

- **Accuracy** for disease classification.
- **AUC Score** to assess the models' ability to distinguish between classes.
- **Statistical Distances**: Kullback-Leibler Divergence and Brier score for probability distributions.
- **Training and Test Times** for computational efficiency.
- **F1-Score** to balance precision and recall.
- **Log Loss** for probability estimates.

V. Implementation

A. Cross-Validation

Cross-validation (CV) is employed to assess the performance of our probabilistic models. The implementation, located in `utils/train.py`, splits the dataset into k folds. For each fold, the model is trained on $k-1$ folds and validated on the remaining fold. This process is repeated k times, ensuring every data point is used for both training and validation, providing a reliable estimate of model generalisation.

B. Discretisation

Discretisation is a key preprocessing step where continuous variables are transformed into discrete bins for probabilistic analysis. The discretisation methods are implemented in `utils/discretise.py`. We use two binning methods:

- **Sturges' Formula:** The number of bins, k, is calculated as:
$$k = \lceil \log_2 n + 1 \rceil$$
where n is the number of data points.
- **Freedman-Diaconis Rule:** This method calculates the number of bins based on the interquartile range (IQR) and the number of data points:
$$\text{bin width} = 2 \times \frac{IQR}{n^{1/3}}$$

The function `discretise` allows for either method to be applied, or for a custom number of bins to be specified. The helper function `assign_bins` assigns data to bins, ensuring the number of bins does not exceed the number of unique values in each feature.

For further details, the reader is referred to the code repository [1].

VI. Results & Evaluation

A. Model Performance Comparison

The performance of Discrete Bayesian Networks (DBNs), Gaussian Bayesian Networks (GBNs), and Gaussian Processes (GPs) is summarised in Table I. With further results being shown in tables II through VII.

TABLE I
MODEL PERFORMANCE COMPARISON.

Model	Acc (%)	F-1	AUC	Brier	Log Loss	Time (s)
DBN	94.2	0.942	0.944	0.0576	2.08	38.4
GBN	50.0	0.663	0.500	0.187	0.558	1.81
GP	64.3	0.728	0.655	0.198	0.564	4.34

B. Visualisation of Results

Confusion matrices in the notebooks [1] show DBN's clear advantage in prediction accuracy. With charts showing the other model types and their performance also present.

C. Summary

DBN outperformed GBN and GP across all metrics, achieving 94.2% accuracy, 0.942 F-1, and 0.944 AUC. GBN had weak performance (AUC 0.500), and GP had moderate accuracy (64.3%), but was computationally costly.

D. Probabilistic Queries

The relevant evidence was queried using a DBN with a TANB structure and Bayesian parameter estimation in `final-queries.ipynb` [1].

- **Query 1**: $P(Group = 0|\text{Visit} = 2, \text{Age} = 88, \text{EDUC} = 14, \text{SES} = 2, \text{MMSE} = 30, \text{CDR} = 0, \text{eTIV} = 2004, \text{nWBV} = 0.681, \text{ASF} = 0.876)$ 99.4% Chance.
- **Query 2**: $P(Group = 1|\text{visit} = 3, \text{Age} = 80, \text{EDUC} = 12, \text{MMSE} = 22, \text{CDR} = 0.5, \text{eTIV} = 1698, \text{nWBV} = 0.701, \text{ASF} = 1.034)$ 90.4% Chance.
- **Query 3**: $P(Status = 0|\text{MDVP:Fo(Hz)} = 197.076, \text{MDVP:Fhi(Hz)} = 206.896, \text{MDVP:Flo(Hz)} = 192.055, \text{MDVP:Jitter(\%)} = 0.00289, \dots)$ 99.9% Chance.
- **Query 4**: $P(Status = 1|\text{MDVP:Fo(Hz)} = 162.568, \text{MDVP:Fhi(Hz)} = 198.346, \text{MDVP:Flo(Hz)} = 77.63, \dots)$ 99.6% Chance.

VII. Discussion

The results highlight the strengths and weaknesses of each model in predicting dementia and Parkinson's disease. Discrete Bayesian Networks (DBNs) were the most effective, offering high accuracy and interpretability, while Gaussian Bayesian Networks (GBNs) struggled with continuous data due to the limitations of Gaussian assumptions. Gaussian Processes (GPs), though flexible in modelling complex relationships, faced high computational costs, emphasising the trade-off between performance and efficiency with large, real-world datasets. Future work could refine GBN and GP performance through better data pre-processing or hybrid approaches.

References

[1] A. Atkinson, "Medical diagnosis using probabilistic ai," 2024. [Online]. Available: https://github.com/alfieatkinson/Medical-Diagnosis-Using-Probabilistic-AI

[2] G. Battineni, F. Amenta, and N. Chintalapudi, "Data for: Machine learning in medicine: Classification and prediction of dementia by support vector machines (svm)," 2019. [Online]. Available: https://doi.org/10.17632/tsy6rbc5d4.1

[3] M. A. Little, P. E. McSharry, E. J. Hunter, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, 2008. [Online]. Available: https://www.kaggle.com/datasets/uciml/parkinsons-disease-data-set

[4] G. Briganti, M. Scutari, and R. J. McNally, "A tutorial on bayesian networks for psychopathology researchers." *Psychological methods*, vol. 28, no. 4, p. 947, 2023.

[5] M. F. Ramoni and P. Sebastiani, "Learning bayesian networks," in *Intelligent Information Technologies: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2008, pp. 315–321.

[6] E. Driver and D. Morrell, "Implementation of continuous bayesian networks using sums of weighted gaussians," *arXiv preprint arXiv:1302.4942*, 2013.

[7] M. Pozzi and A. Der Kiureghian, "Gaussian bayesian network for reliability analysis of a system of bridges," in *Proceedings of the 11th international conf. on structural safety and reliability, New York, United States*, 2013.

[8] K. Tiwari and N. Y. Chong, "Gaussian process," *Multi-robot Exploration for Environmental Monitoring*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:214063181

[9] J. Bernardo, J. Berger, A. Dawid, A. Smith *et al.*, "Regression and classification using gaussian process priors," *Bayesian statistics*, vol. 6, p. 475, 1998.

[10] E. L. Snelson, *Flexible and efficient Gaussian process models for machine learning*. University of London, University College London (United Kingdom), 2008.

[11] S. McLachlan, K. Dube, G. A. Hitman, N. E. Fenton, and E. Kyrimi, "Bayesian networks in healthcare: Distribution by medical condition," *Artificial intelligence in medicine*, vol. 107, p. 101912, 2020.

[12] N. Friedman and I. Nachman, "Gaussian process networks," *arXiv preprint arXiv:1301.3857*, 2013.

Appendix

TABLE II
PREDICTING DEMENTIA USING DISCRETE BAYESIAN NETWORKS (ALL VALUES ROUNDED TO THREE SIGNIFICANT FIGURES)

Model	Acc (%)	Precision	Recall	F1-Score	ROC AUC	Brier	Log Loss	Time (s)
Tree-Augmented Naive Bayes (Bayes)	94.2	0.985	0.904	0.942	0.944	0.0576	2.07	41.6
Tree-Augmented Naive Bayes (Max Likelihood)	66.8	1.00	0.362	0.531	0.681	0.332	11.9	5.60
BDeu Hill Climbing (Bayes)	94.2	0.985	0.904	0.942	0.944	0.0576	2.07	38.7
BDeu Hill Climbing (Max Likelihood)	66.8	1.00	0.362	0.531	0.681	0.332	11.9	5.73
PC Algorithm (Bayes)	94.2	0.985	0.904	0.942	0.944	0.0576	2.07	38.4
PC Algorithm (Max Likelihood)	66.8	1.00	0.362	0.531	0.681	0.332	11.9	6.10

TABLE III
PREDICTING PARKINSON'S DISEASE USING DISCRETE BAYESIAN NETWORKS (ALL VALUES ROUNDED TO THREE SIGNIFICANT FIGURES)

Model	Acc (%)	Precision	Recall	F1-Score	ROC AUC	Brier score	Log Loss	Time (s)
Tree-Augmented Naive Bayes (Bayes)	79.2	0.893	0.648	0.751	0.788	0.209	7.52	48.4
Tree-Augmented Naive Bayes (Max Likelihood)	64.4	0.879	0.314	0.461	0.636	0.356	12.8	7.65
BDeu Hill Climbing (Bayes)	79.2	0.893	0.648	0.751	0.788	0.209	7.52	36.4
BDeu Hill Climbing (Max Likelihood)	64.4	0.879	0.314	0.461	0.636	0.356	12.8	6.56
PC Algorithm (Bayes)	79.2	0.893	0.648	0.751	0.788	0.209	7.52	37.6
PC Algorithm (Max Likelihood)	64.4	0.879	0.314	0.461	0.636	0.356	12.8	7.50

TABLE IV
PREDICTING DEMENTIA USING GAUSSIAN BAYESIAN NETWORKS (ALL VALUES ROUNDED TO THREE SIGNIFICANT FIGURES)

Model	Acc (%)	Precision	Recall	F1-Score	ROC AUC	Brier score	Log Loss	Time (s)
BDeu Hill Climbing	50.0	0.500	1.00	0.663	0.500	0.187	0.560	1.81

TABLE V
PREDICTING PARKINSON'S DISEASE USING GAUSSIAN BAYESIAN NETWORKS (ALL VALUES ROUNDED TO THREE SIGNIFICANT FIGURES)

Model	Acc (%)	Precision	Recall	F1-Score	ROC AUC	Brier score	Log Loss	Time (s)
BDeu Hill Climbing	37.4	0.416	0.712	0.518	0.380	0.398	1.59	8.69

TABLE VI
PREDICTING DEMENTIA USING GAUSSIAN PROCESSES (ALL VALUES ROUNDED TO THREE SIGNIFICANT FIGURES)

Model	Acc (%)	Precision	Recall	F1-Score	ROC AUC	Brier score	Log Loss	Time (s)
RBF Kernel	61.5	0.588	0.782	0.666	0.616	0.237	0.670	8.79
Matern Kernel	54.9	0.530	0.964	0.678	0.572	0.236	0.660	9.47

TABLE VII
PREDICTING PARKINSON'S DISEASE USING GAUSSIAN PROCESSES (ALL VALUES ROUNDED TO THREE SIGNIFICANT FIGURES)

Model	Acc (%)	Precision	Recall	F1-Score	ROC AUC	Brier score	Log Loss	Time (s)
RBF Kernel	64.3	0.588	0.961	0.728	0.655	0.271	1.16	8.57
Matern Kernel	62.4	0.577	0.943	0.716	0.636	0.278	1.19	8.24