Expectation Maximization For Ensemble Structure Learning in the Presence of Latent Variables using Gaussian Mixture Models and Static Bayesian Networks

Lindelani Delisa Dlamini
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Supervised by Prof. Ritesh Ajoodha
1898593@students.wits.ac.za

***Abstract*—This study explores structure learning in Bayesian networks, with an emphasis on handling latent variables. By simulating data reflective of Alzheimer's Disease interactions, we assess different algorithms' ability to uncover hidden patterns. Our approach utilizes Gaussian Mixture Models and statistical measures such as the Bayesian Information Criterion to direct our search for the most accurate model structure. We tested several algorithms, like Tabu search and Bayesian Model Averaging, on datasets of 2000 and 6000 samples. The results reveal that a hybrid approach, on average, was the most successful in approximating the true model, as indicated by low Kullback-Leibler divergence values. Our work highlights the effectiveness of such interpretable methods in tracking the dependency structure between complex variable relationships, even when some of the data is not directly observable.**

***Index Terms*—score-based structure learning, bayesian networks, naive bayes models, tabu search, bayesian model averaging, hillclimb search, parameter estimation, maximum likelihood estiamtion, bayesian information criterion, knowledge discovery**

I. Introduction

In the realm of machine learning, Bayesian networks serve as a powerful tool for modeling probabilistic relationships among variables and inferring dependency structures. However, the challenge intensifies when the dataset includes latent variables, which can be seen as unobserved factors that influence the observed outcomes. Latent variables are prevalent in many complex systems, including biological networks and disease progression models, where not all influencing factors can be measured or are known.

The complexity of uncovering these hidden patterns is well exemplified in the study of Alzheimer's Disease, where early detection and understanding of the disease's progression are critical [1]. Existing models often struggle to capture the nuanced effects of latent variables, leading to a gap in accurately modeling such diseases. Addressing this gap, our study proposes a score-based structure learning approach with static Bayesian networks to better handle latent variables. By generating synthetic data that mirrors the interactions in Alzheimer's Disease, we evaluate the capability of various algorithms to potentially discover and model the I-equivalent structures of the true underlying probabilistic networks. An NP-Hard problem [2], [3]. Through rigorous testing and validation on datasets of different sizes, we aim to identify algorithms that most effectively reveal the latent structures that could be pivotal in understanding and predicting the progression of complex diseases.

The significance of this research lies in its potential to improve the diagnosis and prognosis of diseases by enhancing the accuracy of predictive models. By providing a clearer, Bayesian-founded picture of the underlying probabilistic relationships, healthcare professionals can benefit from more informed decision-making processes, ultimately leading to better patient outcomes.

II. Data Generation and Specifications

We synthesize dataset D to serve as a stand-in for the type of clinical data typically gathered in the study of Alzheimer's Disease (AD). The dataset was specifically structured to capture the discrete nature of information that might be collected through patient questionnaires or forms. This section details the construction of the ground truth model M^*, the logic underpinning its design, and the process of generating and dividing the dataset.

A. Ground Truth Model M^*

Our ground truth model M^*, a Bayesian network, was architected to reflect the intricate probabilistic relationships inherent in AD progression. It consists of a suite of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, each indicative of observable symptoms or latent factors implicated in AD. The architecture of M^* is embodied by a directed acyclic graph (DAG) $G = (V, E)$, where V denotes the variables in $\mathbf{X}$, and E signifies the directed edges that define the conditional dependencies among these variables.
Below is the graphical representation of the ground truth model M^* used in our study:

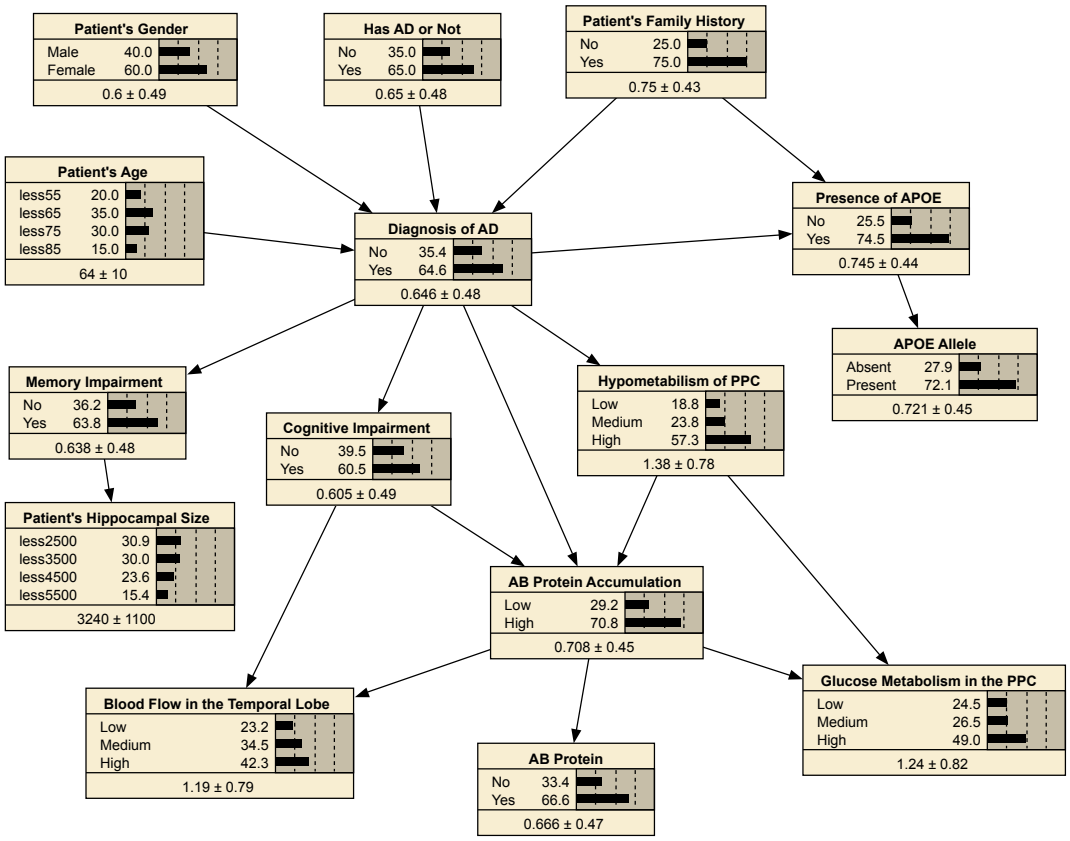

Fig. 1. Ground truth graph model representing the probabilistic dependencies in Alzheimer's Disease progression.

B. Latent Variables Specification

The following latent variables were specified, with their corresponding discrete states, reflecting the nature of clinical data:

- **Latent Variable:** Presence of AD
  Observable Variables: Patient Gender, Family History, Patient Age, AD Test
- **Latent Variable:** AB Protein Levels
  Observable Variables: AB Accumulation, AD Test, Blood Flow
- **Latent Variable:** Presence of APOE Allele
  Observable Variables: Family History, AD Test, APOE, Patient Gender
- **Latent Variable:** Hypometabilism (Metabolic Dysfunction)
  Observable Variables: PPC Glucose Levels, AD Test, Blood Flow
- **Latent Variable:** Cognitive Impairment
  Observable Variables: AD Test, Memory Impairment, Blood Flow, PPC Glucose Levels

C. Division of Dataset

The dataset D was partitioned into a training set Tr, a test set Te, and a validation set Va. The training set Tr is utilized for the initial structure learning. The test set Te is employed to introduce prior knowledge by selectively incorporating random variables with latent information into Tr before the application of Gaussian Mixture Models (GMMs). This process simulates an incremental acquisition of expert knowledge in the structural learning phase. Lastly, the validation set Va is reserved for the assessment of learned models, particularly for computing the Kullback-Leibler (KL) divergence in comparison to M^*.

Our synthetic dataset D, with its comprehensive design and preprocessing, will be fundamental for evaluating the performance of our structure learning algorithms. Providing insights into their ability to potentially uncover complex and hidden patterns within the data.

III. Methodology

Our investigation is aimed at learning model structures M from a dataset D, which is generated from a ground truth model M^*. The models learned are evaluated against M^* using Kullback-Leibler (KL) divergence with evidence Va from the dataset, to facilitate knowledge discovery about the underlying structure of our AD problem. It is crucial that all models M are learned while preserving the state space of

the generated dataset D to ensure the precision of our joint-probability-based KL divergence comparison.

A. Hill-Climbing Search (HCS) & Initial Structures

HCS is a score-based learning approach that incrementally modifies a given structure to better fit the data [4], [5]. This method is particularly advantageous in structure learning due to its ability to incorporate various initial structures. These structures will range from none at all to informed forests based on prior knowledge, such as a set of structures present in M^*, or even structures derived from other algorithms like Chow-Liu trees and Tree-Augmented Naive Bayes (TAN) [6], [7]. The utilization of such informed initial structures can be viewed as a gradual infusion of expert knowledge into the learning process, which is pivotal in accurately capturing the nuances of AD progression. The structure S of a BN is evaluated by a scoring function f on the data D, where the score reflects the fitness of S to D:

$$\text{Score}(S|D) = f(S, D) \tag{1}$$

Our scoring function will use the Bayesian Information Criterion score to balance the likelihood of the data under the structure with a penalty for complexity to prevent overfitting.

B. BIC Score & Mutual Information

For model selection, we use the Bayesian Information Criterion (BIC) score because it is based on the likelihood function and is penalized for the number of parameters to avoid overfitting, which is good for our knowledge discovery task. The BIC score for a model M given data D is expressed as:

$$\text{score}_{\text{BIC}} = \ell\left(\hat{\theta}_{\mathcal{G}} : \mathcal{D}\right) - \frac{\log M}{2}\,\text{DIM}[\mathcal{G}] \tag{2}$$

where M is the number of training instances and the DIM $[\mathcal{G}]$ is the number of independent parameters in the network. For maximum likelihood estimate, $\hat{\theta}$, given a particular graph structure, $\mathcal{G}$, relative to the data, $\mathcal{D}$ [2], [8].

We then use Mutual Information (MI) to rank candidate edges by assessing the information gain between pairs of variables, aiding in the understanding of variable dependencies. The MI for two discrete random variables X and Y is given by:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{3}$$

where $p(x, y)$ is the joint probability mass function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability mass functions of X and Y respectively [9].

C. Expectation-Maximization & Gaussian Mixture Models

The EM algorithm is a powerful tool for the estimation of latent variables within a dataset. When coupled with GMM, it allows for the modeling of complex distributions by assuming that the data is generated from a mixture of several Gaussian distributions. For AD data, which includes a range of biometric and cognitive assessment variables, EM & GMM can facilitate the relearning of our latent variables—such as the underlying genetic predispositions or the presence of biomarkers—that are not directly observed but inferred through observable data. We thus relearn the variables by using this observable data [3], [10], [11].

The EM algorithm iteratively improves the parameters of the GMM by alternating between the expectation (E) step, which estimates the probability of each data point belonging to each cluster, and the maximization (M) step, which computes the parameters that maximize the likelihood of the data given these probabilities:

$$\theta^{(t+1)} = \arg\max_{\theta} \sum_{i=1}^{N}\sum_{k=1}^{K} w_{ik}^{(t)} \log P(x_i|\theta_k) \tag{4}$$

where $\theta^{(t+1)}$ are the parameters at iteration $t+1$, $w_{ik}^{(t)}$ is the probability of data point x_i belonging to cluster k at iteration t, and $P(x_i|\theta_k)$ is the likelihood of x_i given the parameters of cluster k.

D. Tabu Search, Model Averaging, and Ensemble Methods

We supplement Tabu Search [12] with Model Averaging and Ensemble methods, creating a robust framework for structure learning. This combination is particularly effective when integrating Mutual Information for edge ranking and BIC scores for model evaluation using a subset of our training data Tr. In our AD dataset, this method allows us to discern the most informative edges that define the relationships between observed and latent variables. During the structure assembling process we maintain the acyclic nature of the network to ensure that the final assembled structure satisfies the DAG properties essential for a valid BN.

We select edges from the top models and reinforce them by a factor of 2, reflecting their relative importance in the consensus structure:

$$\text{G} = \bigcup_{i=1}^{\min(k,n-1)} \{e_i \mid e_i \in E_{\text{top-ranked}}\} \tag{5}$$

where G is our final acyclic structure $G = (V, E)$, $E_{\text{top-ranked}}$ is the ordered set of edges based on MI scores, and k is the total number of possible edges to consider. This enhanced approach ensures that the final model is both informative and non-overfitting to data.

E. Marginal Likelihood Estimation

We use Bayesian Estimation (BE) for all of our parameter estimation because it is particularly beneficial over the Maximum Likelihood Estimation given how sparse our data is. BE allows for a nuanced trade-off between the evidence provided by the data and our prior beliefs about the disease's progression, resulting in a more reliable estimation of the model parameters [3]. The Bayesian Dirichlet equivalent uniform (BDeu) prior employed in our estimation process, assumes that the prior distribution is equivalent across structures that encode the same set of independencies,

ensuring a uniform treatment of model structures prior to observing the data.

Marginal likelihood integrates out the parameters from the likelihood and prior distribution:

$$P(D) = \int P(D|\theta)P(\theta)d\theta \tag{6}$$

With $P(D|\theta)$ being the likelihood of the data given the parameters, $P(\theta)$ is the prior distribution of the parameters, and the integral sums up (or integrates) over all possible values of θ.

F. KL Divergence

We then use KL Divergence [13], [14] to measure of how our learned model M's probability distribution diverges from our ground truth distribution M*. For discrete probability distributions P (from M^*) and Q (from M), the KL Divergence for a single data point is:

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \cdot \log\left(\frac{P(i)}{Q(i)}\right) \tag{7}$$

For the validation set Va, the average KL Divergence is:

$$\text{Avg } D_{\mathrm{KL}}(P \parallel Q) = \frac{1}{|Va|} \sum_{x \in Va} \sum_i P(x_i) \cdot \log\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{8}$$

.

We calculate these joint probabilities using Variable Elimination, an exact inference algorithm used to compute the marginal distribution over a subset of variables. By evaluating our KL Divergence on the validation set Va, we measure the learned model's generalization to the dataset D produced from the probability distribution P^* induced by the true structure M^* over IID-generated data instances.

IV. Experiment Results

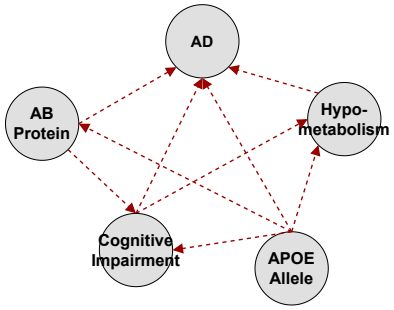

Fig. 2. The dependency structure learned using the Ensemble III algorithm. The shaded nodes represent our latent variables, and the dotted red lines represent the proposed relations between them.

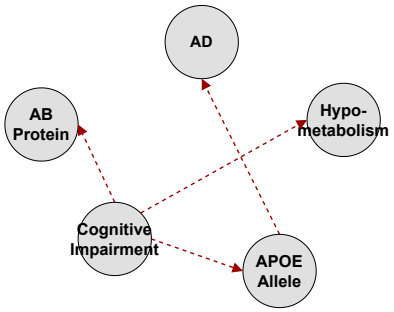

Fig. 3. The graph learned using our Basic HCS algorithm. The shaded nodes represent our latent variables, and the dotted red lines represent the proposed relations between them.

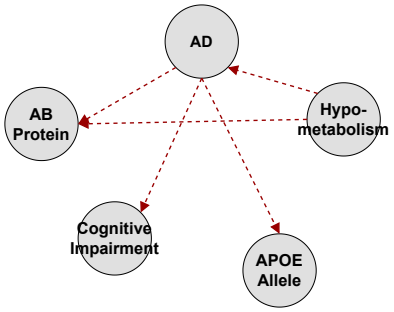

Fig. 4. The influence structure learned using the Tabu Search on HCS algorithm. The shaded nodes represent our latent variables, and the dotted red lines represent the proposed relations between them.

A. Algorithms Review:

The standout performance of Ensemble III (+ Tabu Search + Limited Edges), as indicated by the lowest average KL divergence (0.49), highlights the value of combining comprehensive search strategies with constraints on model complexity. This approach effectively balances in-depth exploration of the network configurations while avoiding the trap of overfitting, which is crucial in the context of AD where the relationships among variables are intricate.

In comparison, algorithms like Basic HCS and Tabu Search on HCS, though effective, demonstrate that the success of an

TABLE I
Summary of Algorithm Performance at 6000 samples

Algorithm	Avg. KL Divergence	Mode Edges
Ensemble III (+ Tabu Search)	0.490001	9
Basic HCS	0.496164	4
Tabu Search on HCS	0.496756	5
Ensemble II (+ BIC Ranking)	0.497739	9
Ensemble (MI Ranking)	0.504949	4
Tabu Search (Limited Edges)	0.507277	4

algorithm also hinges on its ability to explore various network configurations without overlooking potential connections. This balance is particularly important for our neuropsychological dataset, where understanding the nuanced relationships between biological, environmental, and symptomatic factors is key.

It is therefore evident that the complexity of a model does not directly correlate with better performance. Instead, the algorithms that achieve a balance between model simplicity and the ability to capture essential dependencies tend to perform better. The diverse structures and their corresponding performances highlight that AD is a multifactorial condition with complex interactions among its variables, even its unobservable ones. The integrated analysis underscores the potential of using a hybrid approach, combining elements from different algorithms, to develop an optimal model for tracking the influence structure between latent variables.

B. Initial Structures Review:

TABLE II
INITIAL STRUCTURE PERFORMANCE SUMMARY ON 6000 SAMPLE DATASET

Initial Structure	Avg. KL Divergence	Avg. Number of Edges
TAN	0.4913	6.83
Chow-Liu Tree	0.5493	6.24
Random	0.5595	6.74
Informed	0.6297	6.11
None	0.6791	6.20

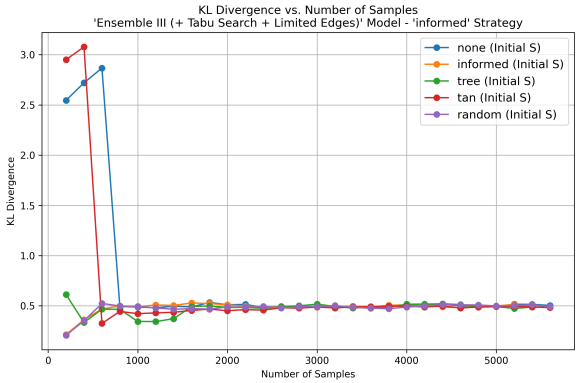

Fig. 5. Comparison of initial structures for Ensemble III at 6000 samples

The performance of the Tree-Augmented Naive Bayes (TAN) initial structure, with the lowest average KL divergence (0.49), shows its effectiveness in capturing the complex relationships in AD data. When comparing the different sample size, it does significantly better with the more data points we get. TAN structures blend simplicity with the ability to represent conditional dependencies, making them well-suited for modeling such diseases like AD.

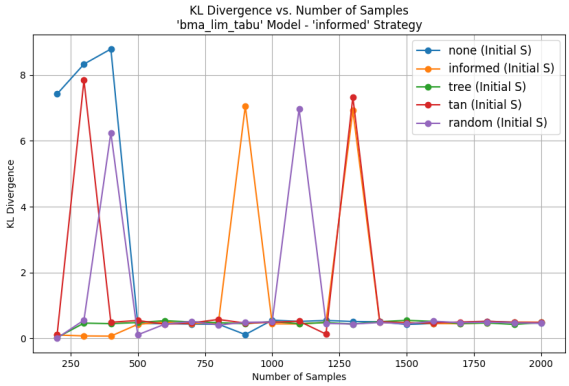

Fig. 6. Comparison of initial structures for Ensemble III at 2000 samples

Chow-Liu Trees and Random structures also provide valuable insights. The tree structures, due to their hierarchical nature, can mirror some of the causal relationships in AD and shows a more stable learned structure when we have smaller data instances. While Random structures introduce an element of unpredictability that can uncover new connections. Conversely, the Informed and None structures show limitations. An informed approach seems to possibly inadvertently add bias to the model, and starting without any predefined structure (none) might lack the necessary direction for efficient learning.

Our experimentation and results show the importance of choosing the right algorithm and right initial structure to accurately capture the probabilistic relationships between latent variables. The insights from this study are crucial for advancing our understanding of modeling such complex interactions and developing more effective knowledge discovery strategies.

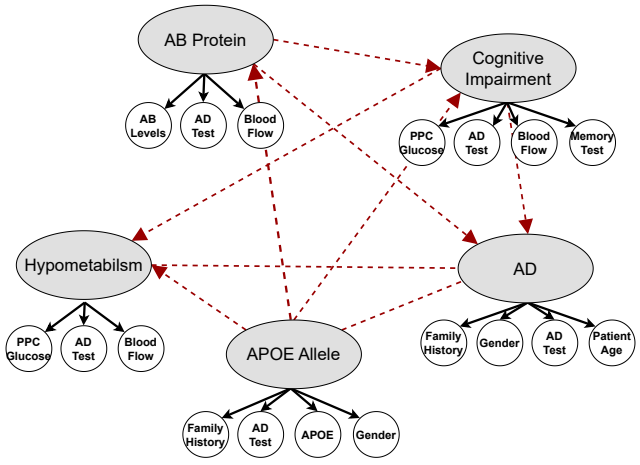

Fig. 7. The full implied structure learned using our Ensemble III algorithm. The shaded nodes represent our latent variables, and the dotted red lines represent the proposed relations between them.

V. Discussion

A. Approach

In our study, we employed heuristic combinatorial optimization methods for structure learning, particularly focusing on addressing the challenge of missing variables. Our experimental results underscore the importance of balancing model complexity against the risk of overfitting, a common pitfall in structure learning.

Our heuristic search strategy is tailored to efficiently explore the model space while escaping local maxima. By utilizing random restarts in the form of random Directed Acyclic Graphs (DAGs), we better learn the desired I-equivalent structures that preserve $M*$'s conditional independence assertions given our lack of need for precise edge orientation. This is seen from how we outperform Tabu Search (Limited) and Ensemble I (MI Rankings), which do not utilize this randomness unlike the variants; Ensemble II and III

The use of TANs and Tree structures for initial models capitalizes on their sparse structures. Such initial structures tend to incorporate significant dependencies while avoiding the complexity that could lead to overfitting, providing a solid foundation for further model refinement. Also, dividing the training data into separate bands and reserving a portion for evaluation of candidate models exploits the strength of BIC in guiding model selection. This, coupled with the perturbation of data, serves to stress-test the models against various data representations, ensuring robustness and reliability.

Our results highlight the effectiveness of these methods in creating models that are well-fitted yet generalizable. The Ensemble III algorithm, in particular, demonstrated commendable performance with smaller datasets, suggesting that our ensemble approach is well-suited to environments where data may be scarce and/or incomplete. It shows promise in avoiding overfitting while capturing the essential structure of the data, which is critical for advancing our understanding of such complex and latent systems.

B. Limitations

Our study's approach to structure learning in Bayesian networks navigates several complexities inherent to modeling conditions like Alzheimer's Disease. Nonetheless, we recognize certain assumptions that stem from our methodological choices.

1) Dependency on Expert Knowledge: The application of our algorithms presumes the availability of domain-specific knowledge, particularly in defining latent variables and initializing the Expectation-Maximization (EM) process with sensible priors. This reliance is mathematically reflected in our model's sensitivity to the initial parameters of the Gaussian Mixture Models (GMMs). The accuracy of the resultant Bayesian network depends on the quality of these expert-provided starting points, which could be challenging to obtain in practice.

2) Assumption of Gaussian Distributions: Our reliance on GMMs for EM inherently assumes that the underlying data distributions are Gaussian or can be adequately modeled as a mixture of Gaussian distributions. This assumption may not hold for all features within the Alzheimer's Disease dataset, where biological data could exhibit skewed or kurtotic distributions, potentially leading to suboptimal parameter estimation and structure learning.

C. Possible Improvements

Reflecting on these limitations, we identify some enhancements to bolster our structure learning approach:

1) Enhanced Global Optimization Techniques: Integrating genetic algorithms, such the Genetic Algorithm (GA) [15] could augment the global optimization of our structure learning. These algorithms, particularly when combined with a tabu list, could widen the search space and potentially yield models that better capture the complex, non-linear interactions within the Alzheimer's Disease dataset. Mathematically, genetic algorithms would introduce a diverse set of candidate solutions, evolving towards an optimal structure through operations akin to biological processes.

2) Alternative to GMMs: To accommodate data that deviates from Gaussian norms, we could explore alternative density estimation techniques. For instance, Kernel Density Estimators or non-parametric Bayesian methods like Dirichlet Process Mixture Models [16] could provide a more flexible mathematical framework to model the underlying distributions of the Alzheimer's Disease biomarkers. We could also tailor the EM algorithm to focus more on, for example; the posterior probabilities of the data instances. Such a modification could be particularly useful for datasets with a high degree of heterogeneity, as is often the case with neuropsychological data.

VI. Conclusion

Structure learning has demonstrated that it is not only possible to identify these hidden patterns but also to do so with a degree of accuracy that supports clinical and research applications. By employing heuristic combinatorial optimization, we have learned that these latent variables can be systematically approached and understood, allowing us to infer possible interactions between them. Notably, the Bayesian approach to computation facilitates the incorporation of prior knowledge into our models, ensuring that the reasoning behind our findings is closely aligned with current scientific understanding. Also, the power of interpretability holds immense value, as it allows for healthcare professionals to make decisions that are informed by a clear rationale [1], [17].

In solving the problem of structure learning in the presence of latent variables given the example use case of AD identification and mapping, we have not only advanced our understanding of this specific condition but have also contributed to techniques that can be employed to unravel similar complexities in other problem domains. Our research stands

as a testament to the power of Explainable AI and Bayesian methods in navigating uncertain terrains and establishes a foundation for future explorations into the vast, uncharted territories of medical science and AI.

VII. Acknowledgement

I would like to thank the various referenced researchers and their work, being in the era of AI development as it improves exponentially, and God for giving me the opportunity and capability to compose this research project. I am grateful.

References

[1] W. Jagust, *Amyloid + activation = Alzheimer's?*, Neuron, vol. 63, no. 2, pp. 141–143, 2009.

[2] Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In Learning from data: Artificial intelligence and statistics V. Springer.

[3] Ajoodha, R., & Rosman, B. (2017). Tracking influence between naïve Bayes models using score-based structure learning. In 2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech) (pp. 122-127). IEEE.

[4] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing Bayesian network structure learning algorithm," *Machine Learning*, vol. 65, 2006.

[5] Chickering, D. M. (2002a). Optimal structure identification with greedy search. Journal of Machine Learning Research, 3(Nov), 507-554.

[6] Chow, C. K.; Liu, C.N. (1968), "Approximating discrete probability distributions with dependence trees", IEEE Transactions on Information Theory, IT-14 (3): 462–467

[7] Friedman N, Geiger D and Goldszmidt M (1997). Bayesian network classifiers. Machine Learning 29: 131–163

[8] R. Ajoodha, "Influence modelling and learning between dynamic Bayesian networks using score-based structure learning," PhD thesis, 2018.

[9] Cover, T. M., & Thomas, J. A. (2006). Elements of information theory. John Wiley & Sons.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1–38, 1977.

[11] G. J. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, New York, 2000.

[12] Glover, F. (1990). Tabu search: A tutorial. Interfaces, 20(4).

[13] S. Kullback and R. A. Leibler, *On information and sufficiency*, Annals of Mathematical Statistics, vol. 22, no. 1, pp. 79–86, 1951.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., John Wiley & Sons, 2006.

[15] Holland, J. H. (1975). "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence." University of Michigan Press, Ann Arbor.

[16] Beal, M. J., & Ghahramani, Z. (2002). The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. Bayesian Statistics, 7, 453-464.

[17] D. Koller and N. Friedman, Probabilistic graphical models: principles and techniques. MIT press, 2009.