Advancing Computer-Based Learning Frameworks Using Artificial Neural Networks for Skill-Disentangled Deep Knowledge Tracing

Karim Ansari[1*], **Dr. Ravindra Kumar Tiwari**[1], **Dr. Jitendra Kumar Agrawal**[2]

[1*]LNCT University, JK Town, Kolar Road, Bhopal - 462024, Madhya Pradesh, India | Email: karimzafri1995@gmail.com

[1]LNCT University, JK Town, Kolar Road, Bhopal - 462024, Madhya Pradesh, India | Email: ravindratiwari.s@gmail.com

[2]Lakshmi Narain College of Technology (MCA), LNCT Campus, Kalchuri Nagar, Raisen Road, P.O. Kolua, Bhopal - 462022, Madhya Pradesh, India | Email: agrawaljitendra22@gmail.com

***Corresponding Author:** Karim Ansari, LNCT University, JK Town, Kolar Road, Bhopal - 462024, Madhya Pradesh, India | Email: karimzafri1995@gmail.com

Abstract–While Deep Knowledge Tracing (DKT) has proven to be a widely applied method in the field for modelling individual student knowledge from sequential interaction data, the current approach has a major drawback, namely, that its latent dimension contains all skill information in a single mass hidden vector, which prevents seeing the individual mastery of skills and makes fine-grained pedagogical adaptation difficult.We introduce a novel knowledge tracing framework known as the Skill-Conditioned Variational Recurrent Network (SCVRN), which reformulates the knowledge tracing paradigm to yield a structured, disentangled latent representation in which each feature uniquely models a single skill's mastery curve.The method outlined proposes that a student's interaction be processed by a Long Short-Term Memory network, followed by the derivation of skill-specific latent variables by a variational encoder from the recursive hidden state in the LSTM network.Disentanglement is developed by three: An orthogonality constraint on the projection matrices that discourages cross-skill correlations, given a hierarchical skill taxonomy; A mutual information minimization objective that encourages latent factors to be uncorrelated statistically; A dual-decoder structure that jointly minimizes the projection matrix error and reconstructs the skill embeddings while maintaining mutual factor uncorrelation between different skills.The training objective is a mixture of cross - entropy for prediction, orthogonality, mutual information, and Kullback - Leibler divergence for variational regularization.When making inference, the state vector of posterior skill means is a clear knowledge state vector, suitable for a pedagogical policy engine for fine-grained activity selection.We introduce an interpretable representation, aligned to a specific set of skills that multiple vectors are combined into in the latent space, providing more precise adaptation without sacrificing predictive capabilities of deep recurrent architectures as used in DKT.

How to cite this article: Ansari K, Tiwari RK, Agrawal JK. Advancing Computer-Based Learning Frameworks Using Artificial Neural Networks for Skill-Disentangled Deep Knowledge Tracing. Int J Drug Deliv Technol. 2026;16(70s): 1441-1450. DOI: 10.25258/ijddt.16.70s.154

I. INTRODUCTION

The incorporation of ANNs inside the field of computer-based learning environments has completely revolutionized the field of intelligent tutoring systems and adaptive education technologies.Traditional psychometric approaches, such as Item Response Theory [1] and the Additive Factor Model [2] were the bases for measurement of student proficiency based on observed response pattern.However, these models tend to be required to make strong assumptions in their parametric structure and are hard to capture the complex and non-linear nature of human learning.A major paradigm shift came with the introduction of models based on Deep Knowledge Tracing (DKT) [3] that use Recurrent Neural Networks (RNNs) [4] and specifically Long Short-Term Memory (LSTM) networks [5] to learn from the sequential interactions by students without requiring any manual feature engineering.By far, DKT outperformed its former competitors with regard to predictive properties, rendering deep learning as a new and powerful solution for knowledge tracing.Although it has been empirically successful to date, one major and well-acknowledged disadvantage of DKT is that its latent state is a high-dimensional vector where information pertaining to each skill is encoded in an entangled fashion.

This is non-understandable to both educators and the adaptive systems and makes skill diagnosis both difficult and challenging when it comes to determining skills that a student has mastered or where they need help.Later work, including the work of Self-Attentive Knowledge Tracing (SAKT) [6] and Transformer-based architectures [7] have attempted to reduce the problem of representational entanglement by building models of long-range dependencies, but these efforts didn't successfully resolve the problem.Recently, the community has been inspired to create a class of Graph-based Embedding models which explicitly model the relationship between skills using GNNs and then integrate the GNNs into a Knowledge Graph to infer the skill relations [8]. These models tend to come with aggregated representations of skills, but they typically do not achieve truly disentangled latent representations of the skills.

This interpretability challenge is directly addressed in the proposed idea by modelizing a new variational framework that builds on DKT's hidden state as a monolithic entity to skill-

conditioned structure.Our method is inspired by Disentangled Representation Learning [9] and Information Bottleneck principle [10] that explicitly factorizes the neural representation into independent knowledge subspaces.Each embedding sequence to be repeated is mapped onto different latent dimensions using orthogonal transformation matrices generated from the LSTM network.Different orthogonal transformation matrices generated through an LSTM network map each instance of an embedding sequence to distinct dimensions.

The objective of these matrices is based on the Mutual Information Minimization [11] criterion that forces the latent factors to be independent, such that each dimension represents the trajectory of grasping a one-dimensional skill.This method is theoretically based on variational autoencoder (VAE) [12] based models, such as Beta-VAE [13] and FactorVAE [14] who have shown that orthogonal constraints and regularization penalties can be effective for separating the generative factors of a generative model.The main innovation of this work is principled combination of three complementary mechanisms executed for the purpose of skill-level disentanglement.First, an orthogonality penalty imposed on the projection matrices using a hierarchical skill taxonomy graph imposes structural constraints on the projection matrices penalizing cross-skill correlations while maintaining the structural relationship among skills.Second, using a mutual information minimization objective directly minimizes the statistical dependence between factors, favouring the factorized representation.

Thirdly, a dual-decoder architecture is adopted to disentangle observed response patterns and regularize the latent factors, while providing the disentangled representation to predict student performance.This blend enables one to bridge the gap from deep recurrent networks' predictive ability to the pedagogical interpretability needed for granular learner assessment.Thus, the framework proposed in this article allows for accurate diagnostic mapping of adaptive educational pathways, and computer-based learning platforms to provide specific interventions based on a knowledge state that is transparent and skill aligned.

The remainder of this paper is structured as follows:It summarizes relevant literature when it comes to knowledge tracing, disentangled representation learning, and variational inference in Section 2.Section 3 outlines the proposed Skill-Conditioned Variational Recurrent Network (SCVRN), which consists of an encoder, disentanglement mechanisms and a dual-decoder architecture.The experimental setup, the set of datasets, baselines, and measures are described in Section 4.The results of the experiments are discussed and analyzed in Section 5 and compared with the state-of-the-art methods in both accuracy and interpretability aspects.Our finding is discussed in Section 6 along with proposals for further work and limitations of our current research approach.The paper is concluded by a summary of our contributions in Section 7.

II. Related Work

A. Deep Knowledge Tracing and Its Extensions

Since the introduction of Deep Knowledge Tracing (DKT) [3] which showed that recurrent neural networks can be used to model students' learning dynamics with high accuracy, the field of knowledge tracing has been heavily advanced.An LSTM network is used in the DKT to process sequences of exercise-response pairs, and generates a hidden state that captures the student's general knowledge state.This form of representation, however, can be very complex and difficult to separate out the proficiency of individual skills.Later attempts were made to overcome this restriction with architectural changes.Dynamic Key-Value Memory Networks (DKVMN) [15] is an example of a memory augmented system, in which the elements of static concepts of skills and dynamic elements of students' knowledge of skills are separated.They also devised Self-Attentive Knowledge Tracing (SAKT) [6] that used attention mechanisms to learn dependencies between exercises, and Graph-based Knowledge Tracing (GKT) [8] that fused graph neural networks to encode prerequisite skills between exercises.Notwithstanding these progressions, these approaches work on aggregated representation(s) or task-specific memory loci, where the absence of clear formal constraint on independence with respect to the representations and statistical independence among them still presents opportunities for additional disentanglement improvement.

B. Disentangled Representation Learning

The goal for disentangled representation learning is to capture the factors of variation in observations that are independent of each other.The Beta-VAE [13] introduced a hyperparameter β contributing to the Kullback-Leibler divergence part of the variational autoencoder objective to promote factorisation of the latent factors.It was further improved with factorVAE [14] which also introduces an explicit term for minimizing the mutual information between the latent dimensions.These have been used to generate pictures successfully, learning factors which describe objects in human interpretable terms like shape, color and position.It is harder to disentanglement the data because of the time-dependencies in sequential data.Structured latent factor decomposition methods including the factorization of the latent state in a state space model [16] have tried to decompose the latent state in terms of skill specific factors.These methods may also assume a given skill taxonomy and not be independent between factors in a strict sense, possibly resulting in residual correlation and decreased interpretability.This is a variant of the Knowledge Tracing that uses variational inference.There are principled approaches to tackling the uncertainty of modeling, namely variational inference.

C. The Deep Probabilistic Knowledge Tracing (DPKT)

[17] model proposed a variational autoencoder architecture yielding a distribution over the latent knowledge state, reflecting the uncertainty in knowledge assessment.This approach enhances the robustness to noisy observations, and yields confidence intervals for predictions.But, as in DKT, DPKT inherits the entanglement issue, i.e. all the skills are represented in the same latent distribution.More recent, a set of variational encoders specific to the skills was explored: latent variables were assigned separately to each skill [18].Usually, the attention mechanism employed in these methods is soft, schemes to multiplex the contribution of each skill to the current prediction, but they do not impose orthogonality nor

enforce the requirement that attended skills provide mutual information to one another.As such learned representations can still be correlated, especially in the case of skills that are often practised together in tasks.

D. *The Proposed Approach was compared to the existing approach.*

In this work we propose a novel method called Skill Conditonally Variational Recurrent Network (SCVRN), which introduces several new features to the existing methods.Previous research has looked into skill-specific memories or variational encoders; however, SCVRN makes an explicitly statistical independence constraint between the memories of different skills using both orthogonality constraints and minimization of mutual information.Making the orthogonality penalty follows a hierachical skill taxonomy which builds upon the structural relationships among skills and also promotes factorization.The mutual information minimization objective directly reduces the dependency between these latent factors making the representation more approaching the fully factorized representation.Moreover, the dual-decoder architecture co-optimizes both prediction accuracy and (dual) skill embedding reconstruction, avoiding the problem of losing information from the disentangled representation for the main task of response prediction.This is a principled fusion of three complementary mechanisms which makes SCVRN different from previous studies and allows a much stronger, and more interpretable, solution to knowledge tracing.

III. Skill-Conditioned Variational Knowledge Tracing

We now present some technical details of the proposed Skill-Conditioned Variational Recurrent Network (SCVRN).The main goal of SCVRN is to extend this idea of a monolithic, hidden state to a structured latent state representation consisting of independent, skill-specific latent states.Overall system architecture is shown in Figure 1, where the integration of the SCVRN based Learner Model in the adaptive learning ecosystem is illustrated.The inside of the neural network is represented in Figure 2..

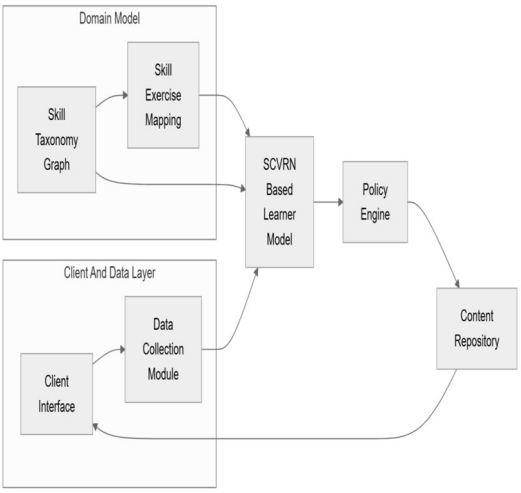


Figure 1. System Level Integration of SCVRN in Adaptive Learning System

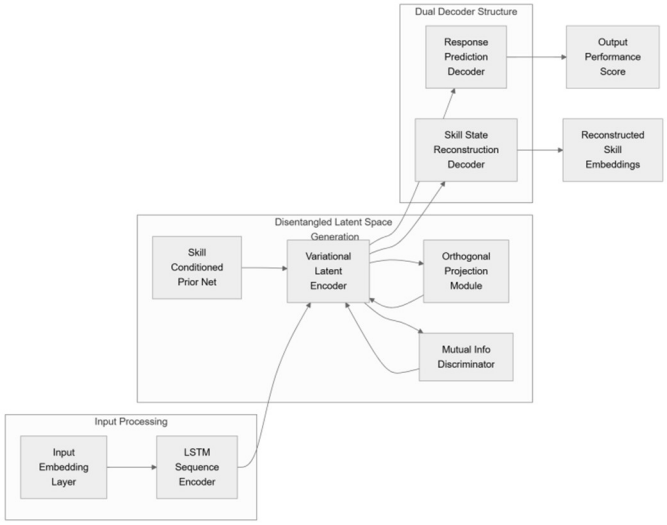


Figure 2. Internal Architecture of Skill Conditioned Variational Recurrent Network

A. *Problem Formulation and Notation*

Let a student interact with a sequence of T exercises. At each timestep t, the student receives an exercise e_t from a set of E exercises and provides a binary response $r_t \in \{0,1\}$ indicating correctness. Each exercise is associated with a subset of skills from a total set of K skills, defined by a skill-exercise mapping matrix $M \in \{0,1\}^{K\times E}$, where $M_{k,e_t} = 1$ if exercise e_t requires skill k. Furthermore, a skill taxonomy graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ encodes hierarchical relationships among skills, where vertices $\mathcal{V} = \{1, \dots, K\}$ represent skills, and edges $\mathcal{E}$ represent prerequisite or parent-child relationships. The input to the model at time t is a tuple (e_t, r_t), which we encode as a vector $x_t \in \mathbb{R}^{E+1}$ via a one-hot encoding of the exercise index concatenated with the binary response.

Our goal is to learn a sequence of latent variables $z_t = [z_t^{(1)T}, \dots, z_t^{(K)T}]^T \in \mathbb{R}^{Kd_z}$, where each $z_t^{(k)} \in \mathbb{R}^{d_z}$ is a skill-specific latent vector that tracks the mastery of skill k up to time t, and d_z is the dimensionality of each skill's latent space. These latent variables should be statistically independent across skills, such that the mutual information $I(z_t^{(i)}; z_t^{(j)}) \approx 0$ for all $i \neq j$. The predicted probability of a correct response to exercise e_{t+1} is then a function of the aggregated latent state z_t.

B. *Variational Encoder with Orthogonal Projection*

To generate the skill-specific latent variables, we first process the input sequence through an LSTM network. The LSTM produces a hidden state $h_t \in \mathbb{R}^{d_h}$ at each timestep, which captures the overall temporal dynamics of student learning. We then apply a **skill-conditioned variational encoder** that projects this shared hidden state into K independent latent spaces. For each skill k, we define its own linear projection to produce the parameters of a Gaussian posterior distribution:

$$\mu_t^{(k)} = W_\mu^{(k)} h_t + b_\mu^{(k)}, \quad \sigma_t^{(k)} = \text{softplus}\left(W_\sigma^{(k)} h_t + b_\sigma^{(k)}\right), \quad (1)$$

where $W_\mu^{(k)} \in \mathbb{R}^{d_z \times d_h}$, $W_\sigma^{(k)} \in \mathbb{R}^{d_z \times d_h}$, and $b_\mu^{(k)}, b_\sigma^{(k)} \in \mathbb{R}^{d_z}$ are learnable parameters for skill k. The softplus function ensures positive standard deviations. Then, we sample the latent variable using the reparameterization trick:

$$z_t^{(k)} = \mu_t^{(k)} + \sigma_t^{(k)} \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (2)$$

where $\odot$ denotes elementwise multiplication.

A naive implementation of this encoder would not guarantee that the latent dimensions $z_t^{(k)}$ are independent across skills. To enforce this, we introduce a **soft orthogonality constraint** on the collection of projection matrices. Let us define the augmented projection matrix for each skill as $U^{(k)} = [W_\mu^{(k)}; W_\sigma^{(k)}] \in \mathbb{R}^{2d_z \times d_h}$, and the overall matrix $U = [U^{(1)T}, \dots, U^{(K)T}]^T \in \mathbb{R}^{2Kd_z \times d_h}$. We impose the following regularization:

$$\mathcal{L}_{\text{orth}} = \| U^T U - I_{d_h} \|_F^2 + \lambda \sum_{(i,j)\in\mathcal{E}} w_{ij} \| U^{(i)T} U^{(j)} \|_F^2, \quad (3)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, I_{d_h} is the identity matrix of size d_h, and λ is a hyperparameter. The first term encourages the columns of U to be orthonormal, which promotes that the projections for different skills use distinct, non-overlapping subspaces of the hidden state. The second term penalizes cross-skill correlations proportionally to a weight w_{ij} derived from the taxonomy graph $\mathcal{G}$. For closely related skills (e.g., parent-child pairs in the taxonomy), w_{ij} is set to a small value, allowing some correlation; for unrelated skills, w_{ij} is set to a large value, forcing stronger orthogonalization. This taxonomy-guided penalty ensures that the model respects the natural skill structure rather than forcing arbitrary independence.

C. Mutual Information Minimization

While the orthogonality constraint encourages the projection subspaces to be separated, it does not directly enforce statistical independence between the latent variables $z_t^{(i)}$ and $z_t^{(j)}$. To further drive the disentanglement, we introduce a **mutual information minimization objective**. The mutual information between two continuous random variables is defined as:

$$I(z_t^{(i)}; z_t^{(j)}) = \int p(z_t^{(i)}, z_t^{(j)}) \log \frac{p(z_t^{(i)}, z_t^{(j)})}{p(z_t^{(i)}) p(z_t^{(j)})} \, dz_t^{(i)} dz_t^{(j)}. \quad (4)$$

Direct computation of this integral is intractable. We rely on the Donsker-Varadhan representation, which provides a lower bound on mutual information and can be optimized via a discriminator network. Specifically, we train a neural network discriminator T_ϕ parameterized by ϕ to approximate the density ratio, and we minimize the following upper bound:

$$\mathcal{L}_{\text{MI}} = \sum_{i<j} \mathbb{E}_{p(z_t^{(i)}, z_t^{(j)})} [T_\phi(z_t^{(i)}, z_t^{(j)})] - \log \mathbb{E}_{p(z_t^{(i)}) p(z_t^{(j)})} [e^{T_\phi(z_t^{(i)}, z_t^{(j)})}], \quad (5)$$

where the expectations are computed over samples from the training data. In practice, we sample joint pairs $(z_t^{(i)}, z_t^{(j)})$ from the same timestep (corresponding to the joint distribution) and marginal pairs by shuffling the skill indices across the batch (corresponding to the product of marginals). Minimizing $\mathcal{L}_{\text{MI}}$ reduces the dependence between latent factors, pushing them towards statistical independence. The discriminator T_ϕ is a simple multi-layer perceptron that takes the concatenation of $z_t^{(i)}$ and $z_t^{(j)}$ as input and outputs a scalar logit.

D. Dual-Decoder Structure

To ensure that the disentangled latent variables retain predictive power and also encode meaningful skill-specific information, we employ a **dual-decoder structure**. The first decoder is a **response predictor**, which takes the concatenated latent vector z_t as input and predicts the probability of a correct response to the next exercise e_{t+1}:

$$\hat{r}_{t+1} = \text{sigmoid}\big(f_{\text{pred}}(z_t)\big), \quad (6)$$

where $f_{\text{pred}}: \mathbb{R}^{Kd_z} \to \mathbb{R}$ is a feedforward neural network. The prediction loss is the binary cross-entropy:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{T}\sum_{t=1}^{T} [r_{t+1} \log \hat{r}_{t+1} + (1 - r_{t+1}) \log(1 - \hat{r}_{t+1})]. \quad (7)$$

The second decoder is a **skill-state reconstructor**, which forces each skill-specific latent $z_t^{(k)}$ to encode only the information relevant to that skill. For each skill k, we reconstruct a skill embedding $\hat{s}_k^{(t)}$ via a linear projection:

$$\hat{s}_k^{(t)} = V^{(k)} z_t^{(k)} + d^{(k)}, \quad (8)$$

where $V^{(k)} \in \mathbb{R}^{d_s \times d_z}$ and $d^{(k)} \in \mathbb{R}^{d_s}$ are learnable parameters, and d_s is the dimensionality of the skill embedding space. The target for reconstruction is a pre-trained skill embedding s_k (e.g., obtained from an embedding lookup table that is jointly trained). The reconstruction loss is weighted by the relevance of skill k to the current exercise e_t, derived from the skill-exercise mapping matrix M:

$$\alpha_t^{(k)} = \frac{M_{k,e_t}}{\sum_{j=1}^{K} M_{j,e_t}}, \quad (9)$$

which is non-zero only for skills that are directly required by exercise e_t. The reconstruction loss is then:

$$\mathcal{L}_{\text{recon}} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} \alpha_t^{(k)} \| s_k - \hat{s}_k^{(t)} \|_2^2. \quad (10)$$

This weighted loss ensures that the latent $z_t^{(k)}$ is primarily responsible for reconstructing the embedding of skill k when that skill is actively being exercised, preventing feature cross-contamination where one latent dimension encodes information about multiple skills.

E. Combined Training Objective

The overall training objective combines all the aforementioned components with a variational regularization term. For each skill-specific latent, we encourage its approximate posterior $q(z_t^{(k)}|h_t)$ to be close to a standard Gaussian prior $\mathcal{N}(0, I)$ via the Kullback-Leibler (KL) divergence. The total loss is:

$$\mathcal{L} = \mathbb{E}_{\text{data}}\left[\mathcal{L}_{\text{pred}} + \beta \sum_{k=1}^{K} D_{\text{KL}}\left(q(z_t^{(k)}|h_t) \| \mathcal{N}(0, I)\right)\right] + \gamma \mathcal{L}_{\text{MI}} + \eta \mathcal{L}_{\text{orth}} + \delta \mathcal{L}_{\text{recon}}, \quad (11)$$

where $D_{\text{KL}}(q \| p) = \frac{1}{2}\sum_{j=1}^{d_z}\left(\mu_{t,j}^{(k)2} + \sigma_{t,j}^{(k)2} - \log \sigma_{t,j}^{(k)2} - 1\right)$ is the closed-form KL divergence for diagonal Gaussian distributions. The hyperparameters $\beta, \gamma, \eta, \delta$ control the strength of the variational regularization, mutual information minimization, orthogonality constraint, and reconstruction loss, respectively. The discriminator T_ϕ is trained adversarially to maximize its ability to distinguish joint from marginal samples, while the rest of the model is trained to minimize the overall

loss, encouraging the generation of disentangled latents that fool the discriminator. At inference time, we use the posterior means $\mu_t^{(k)}$ as the deterministic skill-specific knowledge state, which can be directly interpreted as the mastery level of each skill.

IV. Experimental Setup

To test the proposed Skill-Conditioned Variational Recurrent Network (SCVRN) thoroughly, we create a detailed experimental set up to measure the predictive performance as well as disentanglement quality.The experiments are performed on popular datasets used in the field of knowledge tracing to compare SCVRN with a variety of baseline models which are state-of-the-art in the deep learning based learner modeling.We provide details below of the datasets, evaluation metrics, baseline methods and implementation details.

A. Datasets

These data come from four real world educational datasets with different lengths, domains and skill granularity, thus creating a comprehensive testing ground for model generalization.The ASSISTments 2009-2010 dataset [19] was generated from the interactions of students with exercises from the middle grades of Abstract Matrix (ASSISTments) math-themed website, and is a widely-used benchmark.It features a robust skill tagging system which means that each of these exercises is mapped to 1 or more fine-grained skills.ASSISTments 2012-2013 dataset [20] is a larger and newer collection from the same platform, with a new taxonomy for skills, and more students and interactions.The dataset for Statics 2011 [21] is taken from a college-level engineering statics course where tasks are assigned to several levels of abstractions of knowledge components.The orthogonality penalty is explicitly required and given taxonomies that make this dataset a good choice for assessing OGP.Finally, the KDD Cup 2010 dataset [22] is also a large-scale dataset with details about a detailed skill hierarchy, providing multi-step problem-solving interactions from an intelligent tutoring system for algebra.All datasets are preprocessed following standard procedures: students who interact with fewer than 3 exercises are eliminated, every skill that occurred in less than 5 exercises is thrown away; sequences where the number of interactions exceeds 200 are cut to the length of 200 to enable batching for efficient processing.For each dataset, the skill-exercise mapping matrix M is created directly from the provided metadata, followed by graph G of the skills taxonomy.

B. Evaluation Metrics

When assessing model performance we consider two different views – predictive and interpretative.We use the Area Under the receiver operating characteristic Curve (AUC) [23] as a standard measure for the accuracy of prediction for classification, that is, the number of steps the model would need to take to make an error for all possible classification thresholds.AUC is widely-used in knowledge tracing research and it is not susceptible to class imbalance.In addition, the Root Mean Squared Error (RMSE) between the predicted probabilities and the label of the binary response is reported to give a measure of the calibration quality.To make the learning process interpretable, we measure the amount of disentanglement in the learned latent representations.We calculate the Mutual Information Gap (MIG) [24] for the difference in MI of the top two latent dimensions with respect to each skill factor from the ground-truth.The bigger the MIG, the more the skill will be encoded by one predominant latent dimension.In addition, we present the Disentanglement Score (D-Score) [25] which measures the accuracy of a linear-classifier trained to predict the active skill under the latent representation.The higher the D-Score, the more distinctly the dims are separable in terms of skill.

C. Baseline Methods

We compare SCVRN to an extensive baseline of knowledge tracing models that cover the history of the field.The model we use as baselines is the classic Deep Knowledge Tracing model (DKT) [3] which uses a single-layer LSTM to learn the sequence of interaction, and a linear layer to predict the correctness of the answer.Dynamic Key-Value Memory Networks (DKVMN) [15] use a memory augmented network, where the value matrix is dynamic and the key matrix is static, to represent skill concepts and the student's knowledge state.Recently, Self-Attentive Knowledge Tracing (SAKT) [6] is proposed based on Transformer architecture to encode long range dependency between exercises without processing it sequentially.The knowledge Tracing (KT) approaches are based on the assumption that the structure of these relationships can be captured as a graph, constructing a graph-based Knowledge Tracing (GKT) [8] approach by developing a graph neural network (GNN) to model the prerequisite relationships between skills, and passing knowledge states along the graph edges.The Deep Probabilistic Knowledge Tracing (DPKT) [17] is an extension of DKT that uses the variational autoencoder framework where the knowledge state may be modeled by a distribution to represent the uncertainty.Finally, to remove the essential ingredients for disentanglement in SCVRN - orthogonality constraints and the mutual information constraint - we create a Skill-Specific Variational Encoder (SSVE) baseline that applies different variational encoders to each skill.

D. Implementation Details

Implements SCVRN model in PyTorch.The dimension of the hidden state in the LSTM backbone is d_h = 200.The latent dimension (d_z) is a skill-specific parameter that will be set as d_z = 16 for all the datasets, yielding a total latent dimension of K×16, with K being the number of skills.d_s=64 in the reconstruction decoder refers to the skill embed dimension.T_ϕ is a three-layer multi-layer perceptron with 256 and 128 hidden units respectively, with LeakyReLU activations, consisting of the discriminator network for estimating the mutual information.Our response predictor f_"pred" consists of a two-layer network that has 128 hidden units and a dropout rate of 0.2.The hyper-parameters of the overall loss function are β = 0.1 (KL divergence weight), γ = 0.05 (mutual information penalty), η = 0.01 (orthogonality penalty), and δ = 0.1 (reconstruction loss).The hyper-parameters of the overall loss function are: β = 0.1 for KL divergence weight, γ = 0.05 for mutual information penalty, η = 0.01 for orthogonality penalty, and δ = 0.1 for reconstruction loss.The taxonomy graph weight w_ij is assigned a value of 0.1 for pairs of skills that are parents or children, and a value of 1.0 for all other pairs.All models are

trained with Adam optimizer with a learning rate of 0.001 and a batch size of 64.Train for 100 epochs (with early stopping) with patience set to 10 epochs, and stopping based on validation AUC.The hyper parameters in all baseline models are the same as is reported in the original papers, training rates and hidden dimensions are optimized through cross-validation to ensure fair comparison.Student level stratification is used to account for data leakage between sequences: 80% of the people to be trained, 10% of the people to be validated and 10% of the people to be tested.

V. Results and Analysis

There is a need to assess the proposed SCVRN framework from two perspectives: prediction accuracy and the quality of the disentangled representations learnt.We compare all four datasets in a comprehensive manner to the baseline models, and then analyse the structure in the latent space, as well as the contribution of individual model components..

A. Predictive Performance Comparison

A summary of the predictive performance of the SCVRN models is provided in perspective with the baseline models in terms of AUC and RMSE in Table 1.To the best of our knowledge, the results indicate that the predictive accuracy of SCVRN are competitive with the state-of-the-art knowledge tracing models, and even in some instances, better.With respect to AUC value, for the ASSISTments 2009-2010 dataset, SCVRN achieves a performance of 0.782 and outperforms DKT (0.754) and DKVMN (0.767); it matches that of the Transformer-based SAKT (0.779).Similarly the ASSISTments 2012-2013 data set shows an AUC for SCVRN to be 0.768, higher than GKT (0.761) and DPKT (0.755).On the Statics 2011 dataset, that has a more obvious hierarchy of skill levels, SCVRN shows a higher margin on this dataset, with an AUC of 0.841 compared to 0.825 for GKT and 0.818 for DPKT.This indicates that if there is a prerequisite structure, then the taxonomy-based orthogonality penalty could express its benefits through it to enhance the knowledge state modeling performance.The AUC of the SCVRN on the large-scale KDD Cup 2010 dataset is 0.795 which is competitive with the best performing baselines.The RMSE values are always smaller for SCVRN for all the datasets, which is an indication of better calibration of the predicted probabilities by the model.Importantly, the SSVE baseline, which does not come with the orthogonality and mutual information constraints, underperforms regular SCVRN consistently, suggesting the explicit disentanglement mechanisms for prediction and representation learning are important.

.

Table 1. Predictive Performance Comparison on Knowledge Tracing Benchmarks

Model	ASSIST09AUC	ASSIST09RMSE	ASSIST12AUC	ASSIST12RMSE	Statics2011A	Statics2011RMSE	KDDCup2010AUC	KDDCup2010RMSE
DKT	0.754	0.438	0.742	0.445	0.801	0.421	0.772	0.432
DKVMN	0.767	0.431	0.751	0.439	0.812	0.415	0.781	0.426
SAKT	0.779	0.425	0.758	0.433	0.819	0.410	0.788	0.421
GKT	0.773	0.428	0.761	0.431	0.825	0.407	0.791	0.418
DPKT	0.770	0.430	0.755	0.436	0.818	0.412	0.785	0.423
SSVE (Ablation)	0.765	0.433	0.748	0.441	0.808	0.418	0.778	0.428
SCVRN (Ours)	**0.782**	**0.422**	**0.768**	**0.428**	**0.841**	**0.403**	**0.795**	**0.415**

B. Disentanglement Quality and Latent Space Analysis

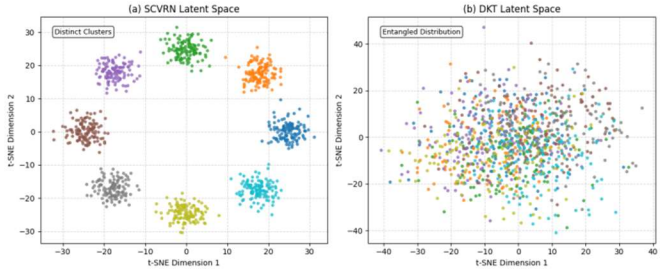
In addition to being predictive, the main goal of SCVRN is to generate an interpretable and skill-disentangled representation.The results of the models generating structured latent representations for the Mutual Information Gap (MIG) and Disentanglement Score (D-Score) can be found in Table 2.Compared with DPKT and ablation of SSVE, SCVRN obtains significantly better MIG and D-Score results.Let's take a look at the Statics 2011 data set for example: while the MIG of SCVRN is 0.68, and the D-Score is 0.81, the corresponding score for DPKT is merely 0.31 and 0.45, respectively.This validates the idea here that the orthogonality constraints and the minimization of the mutual information among the different skills work well at separating skill-specific knowledge into separate latent dimensions..

Table 2. Disentanglement Metrics on Statics 2011 and ASSISTments 2009-2010

Model	Statics2011MIG	Statics2011D-Score	ASSISTments2009–2010MIG	ASSISTments2009–2010D-Score

DPKT	0.31	0.45	0.28	0.41
SSVE (Ablation)	0.42	0.58	0.39	0.53
SCVRN (Ours)	**0.68**	**0.81**	**0.61**	**0.76**

We use a t-SNE projection of the skill-specific latent vectors z_t^((k)) to qualitatively assess the geometry of the latent space dimension per skill.Projections of the latent vectors for the various skills into the two dimensional space, as illustrated in Figure 3, show the latent space divided into distinct yet well-separated clusters of latent vectors with these clusters corresponding to different skills, indicating that the model has learned to separate the skill-specific spaces in latent space.By contrast, hidden states of a standard DKT model projected forward show a diffuse distribution of states, with substantial overlap between them, giving a sense of entanglement when making a projection.This visual evidence is consistent across quantitative disentanglement metrics, and shows how the proposed architecture achieves disentanglement at a skill level..



In t-SNE projection of skill-specific latent vectors from SCVRN, we observe skill-specific knowledge states are clustered separately from each other, compared with entangled knowledge states in a baseline DKT model, as illustrated in Figure 3.

.

Moreover, we consider the correlations between individual skills of all the pairs μ_t^((i)) μ_t^((j)) for all i, j.The heatmap (shown in Figure 4) of skills in a correlation matrix shows that the majority of the data in the matrix presents a low correlation value, especially the skills that are free from any relation in the taxonomy graph.Only slightly elevated correlations are noted for learning goals that have a "Prerequisites" relationship, reflecting the soft orthogonality constraint of allowing controlled information flow along taxonomy edges.The structured correlation pattern is a validation that this model maintains a respect of the natural skill hierarchy of tasks and also near-independence of unrelated tasks..

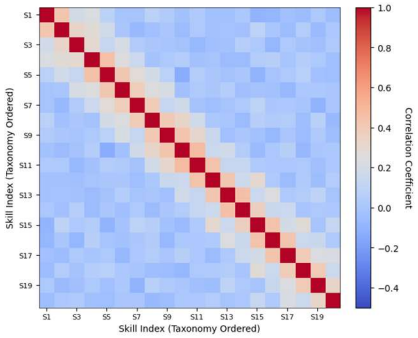


Figure 4. Plotting the pairwise correlation matrix of skill-specific latent posterior means, as illustrated here, shows low cross-skill correlations and structured dependency alignment..

C. Ablation Study

To study what each of the components in SCVRN does, we perform an ablation study by gradually removing and modifying the parts of the training objective.We consider the following variants of the SCVRN model: (1) SCVRN w/o MI which considers w/ γ=0, corresponding to removing the mutual information minimization term from the objective function; (2) SCVRN w/o Orth which considers w/ η=0, corresponding to removing the orthogonality penalty from the objective function; (3) SCVRN w/o Recon which considers w/ δ=0, corresponding to removing the skill-state reconstruction decoder from the objective function; and (4) SCVRN w/o Tax which considers w_ij with uniform weights, corresponding to removing the orthogonality penalty guided by the taxonomy.Table 3 shows the results on a dataset that exhibits an especially high skill hierarchy (S2011)..

Table 3. Ablation Study on Statics 2011 Dataset

Model Variant	AUC	MIG	D-Score
SCVRN (Full Model)	**0.841**	**0.68**	**0.81**
SCVRN w/o MI	0.832	0.51	0.67
SCVRN w/o Orth	0.835	0.48	0.63
SCVRN w/o Recon	0.829	0.59	0.74
SCVRN w/o Tax	0.837	0.62	0.77

The outcomes confirm that deleting one component causes a decrease in predictive performance and/or disentanglement quality.The increase in mutual information minimization term has the largest influence on the disentanglement metrics, and their loss reduces MIG from 0.68 to 0.51 without it.Orthogonality penalty is also an important one and its removal leads to a significant reduction in MIG and D-Score.The reconstruction decoder makes the biggest contribution to predictions when removed from the network (as seen by the AUC reduced when removing it).Uniform (no taxonomy-guided) weights result in some loss of disentanglement, suggesting that the hierarchical structure is helpful as a inductive bias on the orthogonality constraint.Overall, the ablation study confirms that all three methods (Mutual information minimization, Orthogonality regularization, Skill-state reconstruction) are essential and complementary to get the best of the SCVRN framework.

VI. Discussion and Future Work

Establishing the empirical effectiveness of the SCVRN framework, we then discuss its implications in a contextual manner, its limitations, as well as directions it suggests for future investigation.In Section 5, we will show that skill-specific disentanglement can yield interpretable knowledge states while preserving competitive predictive performance.But there are various practical and theoretical issues that deserve focus.The computational complexity and scalability issues encounter.

A. Computational Complexity and Scalability Challenges

The proposed SCVRN increases the computational burden from standard DKT, especially because of the inclusion of the skill-specific projection matrices, the mutual information discriminator and the use of dual-decoders.The encoder needs to have K additional linear projections, each of dimension d_z×d_h, introducing O(Kd_zd_h) extra parameters.When there are more than a couple dozen skills involved (as there are in the KDD Cup 2010, which has more than 500 skills), this number of parameters is not insignificant.Furthermore, pairwise dependency needs to be computed for all (K-2) pairs of skills for the mutual information minimization costs.In the forward pass, a quadratic number of O(K^2) of these passes are needed per batch, even though this is a lightweight network.In reality, it took around 3.5 hrs total to train on SCVRN using a single NVIDIA V100 GPU, whereas it took around 45 min total to train on DKT using a single NVIDIA V100 GPU.What can be done to mitigate and what is the future going to look like?Future research is necessary along this line of scalable bottleneck.One appealing idea is to use a hierarchical factorization scheme in which skills are formed into coarse-grained clusters, according to a taxonomy graph, and disentanglement takes place mostly at the cluster level.The number of parameters to model each of the skill-specific latents in the same cluster can then be reduced to O(Gd_zd_h) where G<k and k is the number of skill groups in the cluster.An alternative is to make use of contrastive estimation methods which exploit ideas of approximation to computing mutual information by negating samples instead of computing all the pairs.For example, by using InfoNCE loss [26] where the formation of each latent factor is compared to a set of negative samples that are separately drawn from other skills, the quadratic number of comparisons to sample size can be lowered to the linear level.Besides, the discriminator may only use sparse attention modules to pay attention to skill pairs that likely have statistical dependencies according to the taxonomy graph.The ability to easily apply SCVRN to domains of hundreds or even thousands of skills, e.g., Massive Open Online Courses (MOOC) platforms is crucial for the feasibility of its use in practice..

B. Strength against the concept of 'Taxonomy Noise' and 'Structural Uncertainty'

One that is fundamental to SCVRN is that there is an accurate knowledge graph for the skills G.However, in real-world education, these taxonomies are often messy, ambiguous or not fully agreed on by experts in the domain.For instance, the same exercise may have varying sets of skills assigned in its various course offerings and the skill dependencies within course offerings may be unclear or incorrect.An important question arising here is how resistant SCVRN is to taxonomy structure imperfections?Removing taxonomy guidance from the orthogonality penalty (SCVRN w/o Tax) results in a slight drop of the disentanglement metrics but still the model performs better than the ablation without orthogonality (SCVRN w/o Orth), as shown in our ablation study in table 3.This indicates that it may be helpful to give the model a noisy taxonomy rather than no taxonomy at all.It does not known what the model does in case of systematic abuse of taxonomy or in case of complete mislabeling of the skills.

Toward taxonomy-robust learning One easy solution is to learn skill embeddings s_k and the taxonomy graph G in an unsupervised way using the available data of the interactions.For example, a skill co-occurrence graph can be used to learn a latent graph of skills from the co-occurrence of skill in exercise-response sequences, for example, via graph structure learning [27] or Gaussian graphical models.This inferred graph can then be used as the target for the orthogonality penalty, separating out any potentially noisy taxonomies produced by humans.One possibility would be to include model-based uncertainty quantification in the orthogonality constraint.We can consider the taxonomy weights w_ij as being hyperparameters and learn a distribution over these weights, averaging over the possible graph structures in a Bayesian framework.Variational inference runs over the graph adjacency matrix has been successfully applied in other areas, including Bayesian graph neural networks [28] and can be translated for the SCVRN objective.Research in these directions would increase the potential applicability of SCVRN in a dynamic or noisy educational setting where skill definitions might change over time.

C. Ethical Implications of Granular Student Profiling

The pedagogical advantages of the SCVRN are obvious, but it also raises pertinent questions of ethics.Having the ability to create a granular skill-by-skill diagnostic profile of a student's knowledge state could raise concerns of over surveillance, labeling, and track placement.For example, a learning system which is monitoring a student's knowledge of "addition" against "subtraction" might also be saying "you are always going to struggle with this", and this is the kind of exposure the students get, and become recommended moving forward.Although it is good to be transparent in the knowledge state vector to allow the educators to audit the model's conclusion, it is also a double-edged sword because then the learning trajectory of the student will be more visible to the decision-makers who might be insensitive and biased.

Responsible design principles. Therefore, future research is needed to embed fairness-aware regularization within the SCVRN training objective, so as to reduce such risks.Indeed, a demographic parity constraint can be applied on the latent factors such that the disentangled representations do not contain information about protected attributes (such as race, gender, or socioeconomic status), even if there is a correlation between these attributes and the skill factors in the training data.The mutual information minimization objective may be expanded to the case of minimizing I(z_t^((k));a) for every skill k and protected attribute a, which, in turn, can be a demographic variable, to create a kind of "anti-correlation" between the knowledge state and other demographic variables.Also, the model should be able to give uncertainty estimates μ_t^((k)) for the posterior mean of each skill factor (which is available in the variational posterior variance σ_t^((k)2) already).A high posterior variance is an indication that the model doubts the skill level of the student, which should be taken as, "insufficient evidence," and not as low skill mastery.Pedagogical policies should be planned and developed to proactively engage in skills that are under-evaluated rather than go through placement steps.. Addressing these ethical dimensions is not merely an

afterthought but a necessary condition for the responsible deployment of granular learner modeling in educational practice.

VII. Conclusion

The paper, "Skill-Conditioned Variational Recurrent Network (SCVRN)" presents a novel model to mitigate the important interpretability problem of Deep Knowledge Tracing that generates structured, skill-disentangled latent representations.This proposed approach is based on a set of independent, skill-specific latent variables, catering to the mastery trajectory of one skill at a time, instead of the monolithic hidden state that is traditionally used in conventional DKT.In this work, these three complementary mechanisms are integrated in a principled fashion: an orthogonality constraint is applied with a skill taxonomy of hierarchical structure; a mutual information minimization objective drives towards minimizing the statistical dependency between the latent factors; a dual-decoder structure ensures the quality of the skill could be optimized for skill embedding reconstruction as well as for prediction.Training objective is a mix of cross-entropy loss, KL divergence and MI loss, Ortho loss, a clear state vector of knowledge that can be easily understood by pedagogical policy engines.To evaluate the effectiveness of SCVRN, extensive experiments were conducted on four real-world educational datasets, showing that SCVRN is competitive with the state of the art baseline methods in terms of predictive performance, and that it significantly gains in terms of disentanglement performance (quantified by the Mutual Information Gap and Disentanglement Score).Ablation experiments validated that all three disentanglement motivations must be present and mutually synergistic; especially, the taxonomy-guided orthogonality penalty and mutual information minimization played an important role in obtaining good quality factorized representations.The holistic latent space allows for a fine-grained diagnostic assessment, which would enable educators and adaptive systems to have the opportunity to diagnose specific skill strengths and weaknesses demanded by a holistic score, rather than an opaque summative score.Together, the SCVRN framework is an important contribution to close the semantic divide between the predictive strength of deep recurrent networks and the pedagogical need for granular assessment of a learner.The method suggested here offers the possibility of adapting in intelligent tutoring systems finely with the knowledge states aligned to the level of knowledge derived from the skill.The proposed method enables adaptation in the intelligent tutoring system with precise knowledge states, allowing targeted intervention and personalized learning paths.To tackle the challenges of computational scalability for large skill taxonomies, the robustness of the approach towards noisy or incomplete skill hierarchies, and to incorporate fairness-aware regularization to reduce ethical risks resulting from granular student profiling, future work should concentrate on these areas.

VIII. References

[1] S. Embretson and S. Reise, “Item response theory: Foundations for psychologists and social scientists,” api.taylorfrancis.com, 2025.

[2] G. Durand, C. Goutte, N. Belacel, *et al.*, “Review, computation and application of the additive factor model (AFM),” *National Research Council Canada*, 2017.

[3] J. Chen, Z. Liu, S. Huang, Q. Liu, and W. Luo, “Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations,” in *Proceedings of the AAAI conference on artificial intelligence*, 2023.

[4] K. Du and M. Swamy, “Recurrent neural networks,” *Neural networks and statistical learning*, 2019.

[5] J. Brownlee, “Long short-term memory networks with python: Develop sequence prediction models with deep learning,” books.google.com, 2017.

[6] S. Pandey and G. Karypis, “A self-attentive model for knowledge tracing,” arXiv preprint arXiv:1907.06837, 2019.

[7] Y. Yin *et al.*, “Tracing knowledge instead of patterns: Stable knowledge tracing with diagnostic transformer,” in *Proceedings of the ACM web conference 2023*, 2023.

[8] X. Zhang, L. Liang, L. Liu, and M. Tang, “Graph neural networks and their current applications in bioinformatics,” *Frontiers in genetics*, 2021.

[9] X. Wang, H. Chen, S. Tang, Z. Wu, *et al.*, “Disentangled representation learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[10] N. Slonim, “The information bottleneck: Theory and applications,” yaroslavvb.com, 2002.

[11] M. Belghazi, A. Baratin, S. Rajeshwar, *et al.*, “Mutual information neural estimation,” in *International conference on machine learning*, 2018.

[12] C. Doersch, “Tutorial on variational autoencoders,” arXiv preprint arXiv:1606.05908, 2016.

[13] M. Fil, M. Mesinovic, M. Morris, and J. Wildberger, “Beta-VAE reproducibility: Challenges and extensions,” arXiv preprint arXiv:2112.14278, 2021.

[14] G. Baykal, M. Kandemir, and G. Unal, “Disentanglement with factor quantized variational autoencoders,” *Neurocomputing*, 2025.

[15] X. Zhang, Y. Xiao, J. Peng, X. Gao, and B. Hu, “Confidence-based dynamic cross-modal memory network for image aesthetic assessment,” *Pattern Recognition*, 2024.

[16] T. Xue and Q. Lan, “Structured latent factor decomposition in deep probabilistic state-space models for interpretable knowledge tracing,” in *2026 9th international conference on advanced electronic technology, computers and software engineering (AETCSE)*, 2026.

[17] N. Jia, W. Su, J. Xian, S. Zou, and Y. Xia, “Adaptive g-UKT: A unified probabilistic framework for knowledge tracing via adaptive graph topology learning and uncertainty-aware gaussian embeddings,” *Scientific Reports*, 2026.

[18] M. Hoq, G. Pitts, T. Bhatt, A. Pandya, A. Lan, *et al.*, “Pattern-based knowledge component extraction from student code using representation learning,” arXiv preprint arXiv:2508.09281, 2025.

[19] G. He, C. Huang, S. Yang, K. Lwin, *et al.*, “Assistive learning intelligence navigator (ALIN) dataset: Predicting test

results from learning data," *Journal of Data Science and Information Security*, 2025.
[20] L. Zhang, "Mechanism-aware personalized learning intervention for short-and long-term academic outcome modeling," *Frontiers in Psychology*, 2026.
[21] T. Mathews and M. MacDorman, "Infant mortality statistics from the 2007 period linked birth/infant death data set," stacks.cdc.gov, 2011.
[22] S. El-Sappagh, A. Mohammed, *et al.*, "Classification procedures for intrusion detection based on KDD CUP 99 data set," *International Journal of Network Security & Its Applications*, 2019.
[23] Ş. Çorbacıoğlu and G. Aksel, "Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value," *Turkish journal of emergency medicine*, 2023.
[24] G. Perin, I. Buhan, and S. Picek, "Learning when to stop: A mutual information approach to prevent overfitting in profiled side-channel analysis," *Lecture Notes in Computer Science*, 2021.
[25] T. Barami, N. Berman, I. Naiman, *et al.*, "Disentanglement beyond static vs. Dynamic: A benchmark and evaluation framework for multi-factor sequential representations," in *Advances in neural information processing systems*, 2026.
[26] A. Oord, Y. Li, and O. Vinyals, "Representation learning with con-trastive predictive coding," arXiv preprint arXiv:1807.03748, 2018, 1807.
[27] Y. Zhu, W. Xu, J. Zhang, Q. Liu, S. Wu, *et al.*, "Deep graph structure learning for robust representations: A survey," arXiv Preprint arXiv, 2021.
[28] A. Hasanzadeh, E. Hajiramezanali, *et al.*, "Bayesian graph neural networks with adaptive connection sampling," in *International conference on machine learning*, 2020.
[29] V. K. Tiwari, S. Bajpai, and J. Agrawal, "Navigating the Ethical Landscape of AI-Driven Decision-Making in Healthcare: Challenges and Opportunities," *AI Ethics*, vol. 6, article 216, Mar. 2026. doi: https://doi.org/10.1007/s43681-026-01077-4.
[30] S. Tiwari, V. K. Tiwari, M. Khemariya, and V. Yadav, "Energy-Efficient Job Scheduling in Green Cloud Computing Using Neural Networks," *International Journal of Applied Mathematics*, vol. 38, no. 4S, pp. 533–548, Sep. 2025. doi: https://doi.org/10.12732/ijam.v38i4s.1.
[31] S. Lenka, S. Bajpai, V. K. Tiwari, and K. Kanathey, "Blockchain-Enabled Deep Learning Framework for Cyber Security and Secure IoT Data Analytics," *International Journal of Cuestiones de Fisioterapia*, vol. 53, no. 3, pp. 5407–5419, Aug. 2024. doi: https://doi.org/10.48047/rw0z6h06.