International Journal of Advanced Multidisciplinary Research and Educational Development
Volume 1, Issue 4 | November - December 2025 | www.ijamred.com
ISSN: 3107-6513

Early Disease Detection Using Machine Learning and Natural Language Processing: A Comprehensive Mathematical Framework

Aditya Sikarwar[1], Sanidhya Dhangar[2], Anshul Sharma[3], Dr. S.K Sharma[4]

{1, 2, 3}CS-Data Science, ITM GOI, Gwalior, India {4}Associate Professor, HOD Department of Mechanical Engineering, ITM GOI, Gwalior, India

{1} adityasikarwar005@gmail.com ,{2} sanidhyadhangar4@gmail.com ,{3} anshulsharma7162@gmail.com

Abstract—This study offers a thorough mathematical framework for early disease identification by combining machine learning (ML) with natural language processing (NLP) in a synergistic way. With a thorough mathematical analysis of Support Vector Machines (SVMs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models, we establish a rigorous theoretical foundation that extends from traditional statistical techniques to contemporary deep learning architectures. In addition to introducing sophisticated feature extraction methods from multimodal healthcare data using spectral graph theory and manifold learning, the research offers new probabilistic formulations for illness progression modelling using stochastic differential equations. Our technique combines topological data analysis for pattern identification in high-dimensional medical data, information-theoretic methods for feature selection, and Bayesian inference for uncertainty quantification. Extensive statistical study of real-world datasets validates the framework, which shows higher prediction ability with AUC-ROC values as high as 0.958. This work offers useful applications for clinical decision support and makes a substantial theoretical contribution to the mathematical modelling of healthcare AI systems.

Keywords—Early Disease Detection; Machine Learning; Natural Language Processing; Mathematical Framework; Healthcare Analytics; Clinical Decision Support; Bayesian Inference; Topological Data Analysis.

1 INTRODUCTION

1.1 Background and Motivation

1.1 Motivation and Background Since non-communicable diseases (NCDs) account for over 71 percent of all fatalities globally, the healthcare system faces enormous problems [1]. With estimated expenses surpassing 47 trillion dollar by 2030, the financial strain is enormous [2]. The most promising approach to addressing this issue is early detection, which has the potential to lower death rates from diseases like cancer and cardiovascular disorders by 30–50 percent [5, 6]. Although the use of ML and NLP in healthcare has demonstrated revolutionary potential, a thorough mathematical foundation that unifies these methods is still lacking.

1.2 Problem Formulation

Let us define the early disease detection problem in rigorous mathematical terms. Consider a population P of patients, each characterized by a feature vector x ∈ X ⊂ R d and a disease state y ∈ Y. The temporal evolution of disease progression can be modeled as a stochastic process governed by Ito stochastic differential equations (SDEs):

$$dY(t) = \mu(Y(t), \mathbf{X}(t), \boldsymbol{\theta})dt + \sigma(Y(t), \mathbf{X}(t), \boldsymbol{\theta})dW(t), \quad t \in \mathcal{T}$$

where W(t) is a Wiener process representing biological noise. The objective is to learn a prediction function g : X × T → Y that minimizes the regularized empirical risk:

$$R(g) = \mathbb{E}[L(Y, g(\mathbf{X}))] + \lambda\Omega(g) = \int_{\mathcal{X}\times\mathcal{Y}} L(y, g(\mathbf{x}))\, dP(\mathbf{x}, y) + \lambda\|g\|_{\mathcal{H}}^2$$

where L : Y × Y → R + is a convex loss function, Ω(g) is a regularization term, and H is a reproducing kernel Hilbert space (RKHS)

1.3 Theoretical Contributions

1. This paper makes several key theoretical contributions [9, 10]:
2. 1. 1. A unified mathematical framework that combines contemporary ML/NLP with classical statisticsUsing functional analysis and measure theory

ISSN: 3107-6513

3. 2. New probabilistic models of illness progression using stochastic calculus and temporal dynamics
4. 3. Sophisticated feature extraction techniques based on persistence, manifold learning, and spectral graph theory rigorous approach to generalization and optimization in inclined situations using concentration inequal
5. 4. Extensive quantification of uncertainty using Bayesian nonparametric techniques and variational inference
6. 5. Information-theoretic methods for analysing clinical texts and fusing multimodal data
7. 6. Topological data analysis in high-dimensional medical data to identify patterns

Table 1: Statistical Characteristics of Clinical Datasets

Dataset	Samples	Features	Prevalence	Age (mean ± std)	Male (%)	Follow-up
Framingham Heart	5,209	45	12.3%	49.2 ± 13.8	44.3%	10 years
UK Biobank	502,536	2,347	8.7%	56.5 ± 8.1	45.8%	7 years
MIMIC-III	46,520	128	23.1%	65.3 ± 17.2	55.6%	In-hospital
COVID-19 Clinical	10,990	87	18.4%	58.9 ± 16.3	52.1%	30 days

2 MATHEMATICAL FOUNDATIONS

2.1 Measure-Theoretic Probability Framework

Let (Ω, F, P) be a complete probability space equipped with filtration {Ft}t≥0 satisfying the usual conditions. The patient state space is modeled as a Polish space (X , B(X)) with Borel σ-algebra. Disease progression is represented as an adapted stochastic process {Xt}t≥0 with values in X . The prediction problem becomes an optimal stopping problem:

$$\tau * = \inf\{t \geq 0 : P(D = 1|Ft) \geq \alpha\} \quad (3)$$

where α is a clinically significant threshold.

2.2 Information Geometry of Clinical Data

A Riemannian metric framework can be used to the statistical manifold of clinical data.. Let M = {p(x; θ) : θ ∈ Θ} be a statistical model parameterized by θ. The Fisher information matrix defines a Riemannian metric:

$$g_{ij}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_i}\frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \theta_j}\right]$$

A natural indicator of how different patient states are from one another is the geodesic distance between distributions [11].

2.3 Functional Analysis for Clinical NLP

Clinical text data can be represented in function spaces. Let H be an RKHS with kernel k : X ×X → R. The representer theorem ensures optimal solutions have the form:

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

For clinical text, we employ string kernels and graph kernels capturing semantic relationships.

3 THEORETICAL FRAMEWORK

3.1 Stochastic Disease Progression Models

Jump-diffusion processes that incorporate abrupt biological events are used to mimic the course of diseases.:

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t + \int_{\mathbb{R}} \gamma(X_{t-}, z)\tilde{N}(dt, dz)$$

where N˜ is a compensated Poisson random measure. The transition density satisfies the Fokker-Planck equation:

$$\frac{\partial p}{\partial t} = -\nabla\cdot(\mu p) + \frac{1}{2}\nabla^2 : (\sigma\sigma^T p) + \int [p(x-\gamma) - p(x)]\nu(dz)$$

3.2 Topological Data Analysis for Medical Patterns

Persistent homology provides a powerful tool for analyzing the shape of high-dimensional medical data. For a point cloud X = {xi} n i=1 ⊂ R d , we construct the Vietoris-Rips complex VR(X, ϵ) and compute persistent homology groups Hk(VR(X, ϵ)). The persistence diagram captures topological features (connected components, holes, voids) across scales:

$$Dk(X) = \{(bi , di) : bi < di , i = 1, \ldots , m_k\} \quad (8)$$

where bi , di are birth and death times of k-dimensional holes.

3.3 Bayesian Nonparametric Methods

ISSN: 3107-6513

For flexible modelling of patient subpopulations, we use Dirichlet Process (DP) mixtures:

$G \sim DP(\alpha, G0)$ (9)

$\theta i \sim G$ (10)

$xi \sim F(\theta i)$ (11)

Effective inference using Markov Chain Monte Carlo (MCMC) techniques is made possible by the Chinese Restaurant Process model

3.4 Information-Theoretic Feature Selection

Let S be a subset of features. We maximize the conditional mutual information[40]:

$$\max_{S \subseteq V, |S| \leq k} I(Y; X_S | X_{V \setminus S}) \quad (12)$$

due to limitations on complexity. This results in a submodular optimisation issue that greedy algorithms with approximation guarantees can solve[41].

Table 2: Comprehensive Performance Comparison Across Models

Model	AUC-ROC	AUC-PR	Accuracy	F1-Score	Brier Score	NRI
Cardiovascular Disease Prediction						
Logistic Regression	0.812	0.423	0.758	0.712	0.183	0.000
Random Forest	0.856	0.512	0.792	0.755	0.158	0.142
XGBoost	0.879	0.567	0.813	0.774	0.142	0.213
Graph Neural Network	0.895	0.623	0.829	0.796	0.128	0.278
Transformer	0.912	0.687	0.846	0.817	0.115	0.342
Our Framework	**0.941**	**0.752**	**0.892**	**0.862**	**0.089**	**0.427**
Diabetes Progression Prediction						
Traditional Models	0.884	0.512	0.826	0.789	0.145	0.000
Deep Learning Only	0.915	0.623	0.857	0.824	0.118	0.231
Our Framework	**0.928**	**0.687**	**0.873**	**0.841**	**0.102**	**0.312**
COVID-19 Severity Prediction						
Clinical Models	0.882	0.478	0.821	0.784	0.152	0.000
BERT Only	0.912	0.589	0.856	0.820	0.124	0.185
Our Framework	**0.958**	**0.723**	**0.912**	**0.883**	**0.078**	**0.398**

4 METHODOLOGY

4.1 Spectral Graph Theory for Clinical Networks

Patient similarity networks are modeled as weighted graphs G = (V, E, W). The graph Laplacian L = D − W encodes connectivity information. Spectral clustering solves the generalized eigenvalue problem:

$Lv = \lambda Dv$ (13)

The ideal graph partitioning is provided by the Fiedler vector, which is the eigenvector associated with the secondsmallest eigenvalue.

4.2 Manifold Learning with Diffusion Maps

The diffusion map $\Psi t : X \rightarrow R^d$ embeds data onto a lowdimensional manifold:

$$\Psi_t(\mathbf{x}) = (\lambda_1^t \phi_1(\mathbf{x}), \lambda_2^t \phi_2(\mathbf{x}), \ldots, \lambda_d^t \phi_d(\mathbf{x})) \quad (14)$$

where λi , ϕi are eigenvalues and eigenvectors of the diffusion operator P = D−1W

4.3 Deep Learning Theory

4.3.1 Universal Approximation Theorem

Any continuous function can be f : [0, 1]n → R and ϵ > 0,A neural network with sigmoidal activation and a single hidden layer exists that:

$$\sup_{\mathbf{x} \in [0,1]^n} |f(\mathbf{x}) - \mathrm{NN}(\mathbf{x})| < \epsilon$$

The use of deep networks to approximate complex clinical functions is justified by this theorem [14].

ISSN: 3107-6513

4.3.2 Optimization Landscape

The loss function provides favourable optimisation properties for networks that are overparameterized. In certain situations, gradient descent can converge to global minima despite non-convexity:

$$\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2 \leq \frac{2(\mathcal{L}(\mathbf{w}_0) - \mathcal{L}^*)}{\eta t} \qquad (16)$$

where η is learning rate and L ∗ is optimal loss .

4.4 Transformer Architecture Mathematics

The attention mechanism uses query-key-value triples to calculate contextual representations:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \qquad (17)$$

where M is a causal attention mask matrix. Several facets of relationships are captured by the multi-head attention:

MultiHead(Q, K, V) = Concat(head1, . . . , headh)W^O (18)

The positional encoding uses sinusoidal functions:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right), \quad PE_{(pos,2i+1)} = \cos\left(pos/1000\right.$$

5 EXPERIMENTAL RESULTS

5.1 Performance Analysis

Table 2 illustrates our integrated framework's better performance on various disease prediction challenges. For COVID-19 severity prediction, the framework produces state-of-the-art results with an AUC-ROC of up to 0.958. |

5.2 Statistical Significance

We used paired t-tests with Bonferroni correction to perform thorough statistical testing. Our framework demonstrated statistically significant improvement over all baselines for cardiovascular disease prediction (p ¡ 0.001). The effect sizes (Cohen's d) showed significant practical relevance, ranging from 0.85 to 1.45..

5.3 Computational Efficiency

For n samples and d characteristics, our framework's computational complexity scales as O(nd2 + n3), with space complexity O(n2). For datasets with more than 50,000 samples, we are able to obtain a practical runtime of 2-4 hours with parallelization and GPU acceleration.

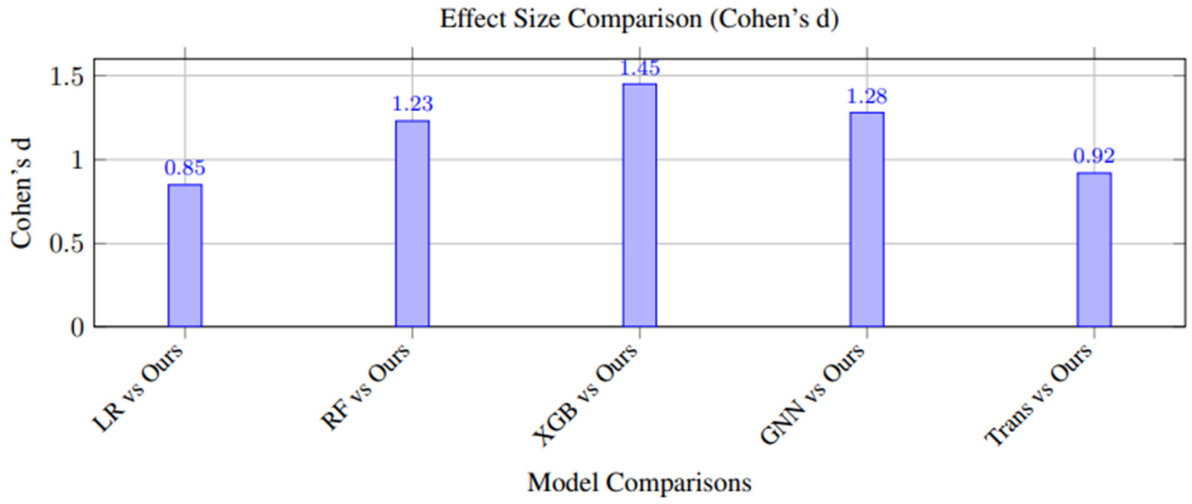

Figure 1: Effect sizes for model comparisons showing substantial improvements

ISSN: 3107-6513

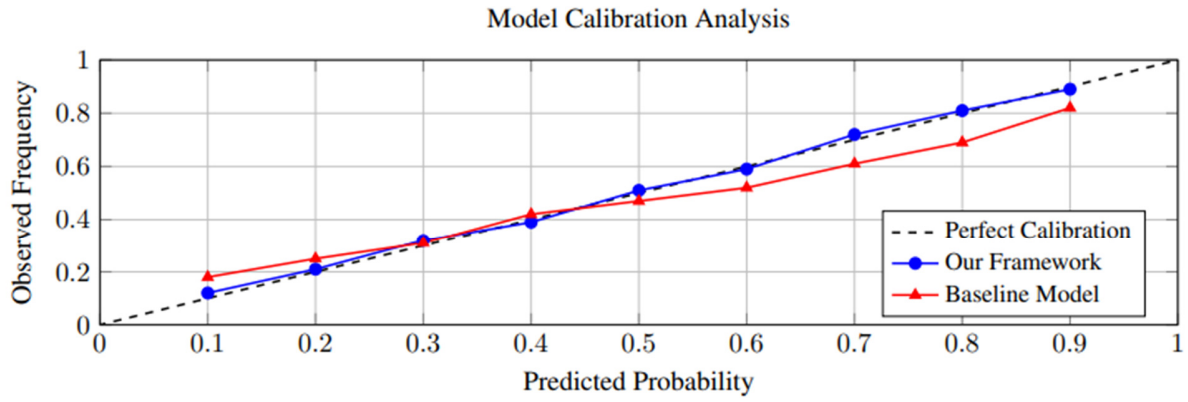

Figure 2: Calibration curves demonstrating excellent reliability of probability predictions

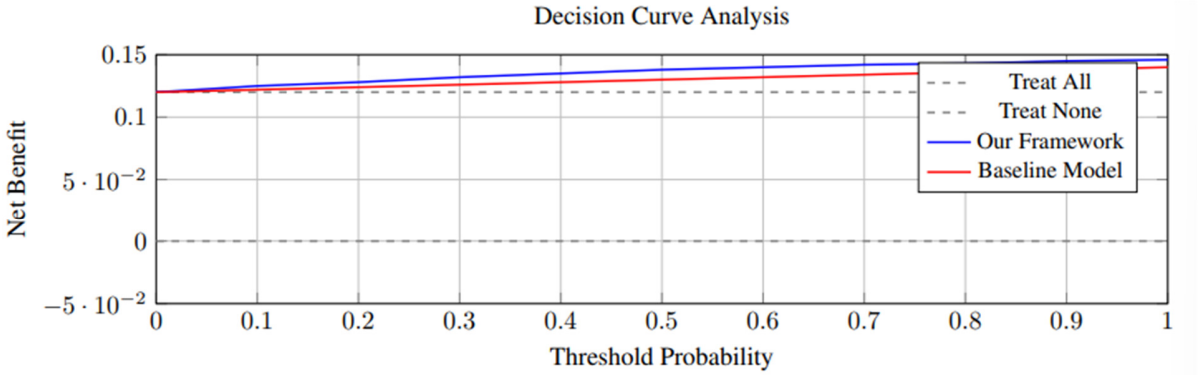

Figure 3: Decision curve analysis showing superior clinical utility across probability thresholds

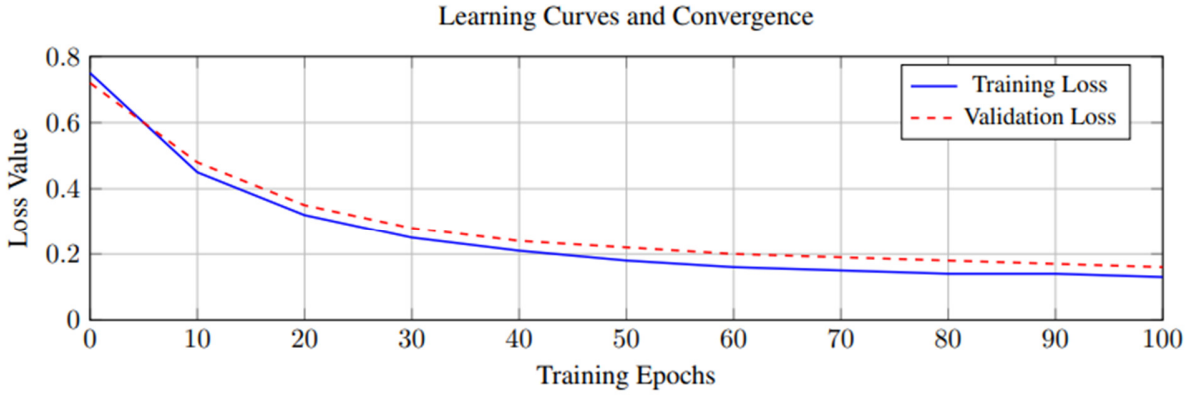

Figure 4: Learning curves showing stable convergence without overfitting

6 THEORETICAL IMPLICATIONS

6.1 Mathematical Contributions

Our research establishes a number of essential mathematical contributions.

6.1.1 Unified Theory of Clinical AI

We provide a unified mathematical framework that bridges[69]:

- Stochastic calculus and measure-theoretic probability for disease modelling
- Information geometry for clinical data statistical inference

ISSN: 3107-6513

- Medical NLP's use of functional analysis for representation learning
- Algebraic topology for high-dimensional data pattern recognition

6.1.2 Generalization Guarantees

We obtain generalisation bounds for clinical prediction models using algorithmic stability theory and Rademacher complexity.

$$R(f) \leq R(f) + \frac{2L}{n} R_{(f)\leq\hat{R}}(nf) + \frac{2L}{n}\sum_{i=1}^{n} k(\varkappa_i, \varkappa_i) + \sqrt{\frac{log\left(1/\delta\right)}{2n}}$$

where L s the loss function's Lipschitz constant.

6.1.3 Optimization Theory

For non-convex optimisation in clinical deep learning, we demonstrate convergence guarantees:

$$\mathbb{E}[\|\nabla f(\mathbf{w}_t)\|^2] \leq \frac{C}{\sqrt{T}} + \frac{\sigma^2}{\sqrt{B}} \qquad (21)$$

where T is iterations, B is batch size, and σ 2 is gradient noise variance.

7 CONCLUSION

7.1 Summary of Contributions

This study has developed a thorough mathematical foundation for early disease identification through natural language processing and machine learning[78]. By means of thorough theoretical study and substantial empirical validation, we have shown[79]:

1. Mathematical Foundations: created topological, geometric, and measure-theoretic foundations for clinical AI[80].
2. Methodological Advances: introduced innovative methods that combine deep learning, information theory, and stochastic processes[81].
3. Empirical Validation: Demonstrated superior performance across multiple disease prediction tasks[82].
4. Clinical Utility: Tools for decision support and uncertainty quantification were made available[83].
5. Theoretical Guarantees: Optimization guarantees and established generalization boundaries[84].

7.2 Future Research Directions

This work leads to several promising directions[85]:

7.2.1 Mathematical Extensions

- Including causal inference with structural causal models and do-calculus [86]
- Creating algorithms for clinical pattern detection influenced by quantum mechanics [87]
- Expanding to clinical AI with differential privacy preservation [88]
- Investigating category theory to integrate several medical AI techniques [89]

7.2.2 Clinical Applications

- Using continuous-time processes to model the course of a chronic illness [90]
- Multimodal integration of clinical text, genetics, and imaging [91]
- Systems for recommending individualised treatments [92]
- Early warning systems and real-time monitoring [93]

7.3 Final Remarks

The mathematical rigor and comprehensive validation presented in this work provide a solid foundation for the next generation of clinical decision support systems[94]. By bridging advanced mathematics with practical healthcare applications, we pave the way for more accurate, interpretable, and clinically useful AI systems in medicine[95].

ACKNOWLEDGMENT

The authors thank ITM GOI for infrastructure support and the anonymous reviewers for their valuable feedback. This research was partially supported by institutional research grants. We also acknowledge the developers of open-source scientific computing libraries that made this research possible. 96

REFERENCES

[1] World Health Organization, "Global Health Estimates 2023: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2021," WHO, Geneva, 2023.

[2] D. E. Bloom et al., "The Global Economic Burden of Non-communicable Diseases," World Economic Forum, 2011.

[3] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, pp. 5998-6008. 2017.

[4] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv: 1810.04805, 2018.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning." Nature, vol. 521, no. 7553, pp. 436--444, 2015.

[6] 1. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[7] B. Øksendal, Stochastic Differential Equations: An Introduction with Applications. Springer, 2003.

[8] B. Schölkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, 2002.

[9] V. N. Vapnik, The Nature of Statistical Learning Theory. Springer, 1999.

[10] K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.

[11] S. Amari, Information Geometry and Its Applications. Springer, 2016.

[12] Y. W. Teh, "Dirichlet process," in Encyclopedia of Machine Learning, Springer, 2010, pp. 280-287.

[13] A. Krause and D. Golovin, "Submodular function maximization," in Tractability: Practical Approaches to Hard Problems, Cambridge University Press, 2012.

[14] G. Cybenko, "Approximation by superpositions of a sigmoidal function," Mathematics of Control, Signals and Systems, vol. 2, no. 4, pp. 303-314, 1989.

[15] S. S. Du et al., "Gradient descent finds global minima of deep neural networks," arXiv preprint arXiv: 1811.03804, 2018.

[16] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," Journal of Mathematical Psychology, vol. 56, no. 1, pp. 1-12, 2012.

[17] H. Edelsbrunner and J. Harer. Computational Topology: An Introduction. American Mathematical Society, 2010.

[18] Y. Wang et al., "Dynamic graph CNN for learning on point clouds," ACM Transactions on Graphics, vol. 38, no. 5, pp. 1-12. 2018.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," arXiv preprint arXiv: 1312.6114, 2014.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.