UNIVERSITY OF NEW BRUNSWICK

FREDERICTON

FACULTY OF COMPUTER SCIENCE

CS6735: MACHINE LEARNING AND DATA MINING

RESEARCH PROJECT REPORT

GROUP MEMBERS

PRIYANKA BHAMARE (3741282)

RISHABH KALAI (3740704)

SADHANA SURESH CHETTIAR (3733403)

Table of Contents

1. Motivation of ML in Healthcare

Machine Learning (ML) is transforming the healthcare sector, influencing medical research, diagnosis, and treatment. It is immensely impacting areas like medical research, diagnosis, and treatment. By analyzing large datasets, ML algorithms can classify and make predictions in complex biological systems. This leads to more accurate diagnostic tools, faster development of new treatments, and personalized medicine.

A great example is how ML helped manage pandemics like COVID-19. The crisis showed we needed quick, data-driven responses to big public health challenges. Technologies like Bayesian Belief Networks were key in limiting disease spread, forecasting case numbers, and guiding policies. Developing mRNA vaccines faster with ML shows we can respond better to global health issues.

The expanding influence of machine learning (ML) in healthcare, notably in the analysis of genetic data, is revolutionary. ML's ability to dissect intricate genetic information is opening doors to fresh avenues for diagnosing diseases, enriching treatment choices, and enhancing patient results. By pinpointing genetic markers and patterns, ML enables personalized medicine strategies and customizing treatments based on individual genetic compositions.

Additionally, ML is revolutionizing drug discovery and development. This discipline used to include research methods that were time-consuming and expensive, but techniques like generative models and reinforcement learning are streamlining it. By efficiently searching large chemical spaces and predicting molecule activity, ML is speeding up new therapy discoveries and optimizing them for clinical use.

Integrating ML into healthcare is more than just tech progress - it is a big shift towards precise, efficient, and personalized care in the future. As we tap more into ML's potential, we are on the edge of a new era in healthcare where data insights pave the way for ground-breaking advances in disease management and treatments.

2. Pandemic Management Through Machine Learning

The narrative surrounding Artificial Intelligence (AI) has undergone a profound transformation, evolving from speculative discussions within the domain of science fiction to playing a pivotal role in addressing some of the most pressing challenges in modern medicine and public health. The advent of generative AI technologies, such as ChatGPT in late 2022, marked a significant leap in AI's emergence into public and academic discourse, demonstrating complex problem-solving capabilities and engaging dialogues. Despite this recent surge in visibility, AI's deployment in critical sectors, notably in the swift development of mRNA vaccines against COVID-19, exemplifies its foundational impact on global health outcomes long before.

At Moderna, AI was integral to the conceptualization and realization of their highly effective mRNA vaccine, enhancing the company's research and development capabilities and enabling a significant increase in mRNA production. This facilitated the generation of optimized mRNA sequences for heightened protein synthesis within the human body, exemplifying a paradigm shift towards a dynamic, responsive approach to health crises. This report delves into the integration of Bayesian Belief Networks (BBNs) in pandemic management, focusing on the transformative role of AI in expediting the development and efficacy of mRNA vaccine technologies. It explores AI's contributions to mRNA methods, underscoring the necessity of leveraging digital innovations for agile, effective responses to pandemics, within the context of evolving public health strategies. Some useful methods through Machine Learning (ML) include multivariate linear regression for case number prediction, exponential smoothing for historical data analysis, and employing compartmental models like the SEIRD for epidemic dynamics prediction [1]. This research emphasizes BBN as a primary focus, providing a detailed exposition on its working and application, reflecting the complexity and urgency of navigating global health challenges with advanced technological interventions.

2.1 Bayesian Belief Networks

The literature review in [2] outlines the significance of early detection and management in the context of novel diseases, highlighting the challenges faced by healthcare systems worldwide during the COVID-19 pandemic. It discusses the evolution of predictive models and their applications in epidemiology, particularly focusing on Bayesian belief networks (BBNs) for their probabilistic and interpretable nature, which allows for better decision support under uncertainty.

2.1.1 Definition & Working

A Bayesian Belief Network is a graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). In simpler terms, a BBN is a diagram that displays how several factors or variables influence one another, incorporating the likelihood of various outcomes based on these relationships.

Figure 1 illustrates a Bayesian Belief Network's structure, showing the relationships between four variables in a Directed Acyclic Graph (DAG). Here is an explanation, referencing the diagram:

- Nodes and Variables: Nodes 1 to 4 represent variables that may correspond to real-world entities or conditions. For instance, Node 1 could represent a medical symptom or an environmental factor.
- Edges and Dependencies: The arrows indicate causal or probabilistic influences. For example, Node 1 (Variable 1) influences both Node 2 (Variable 2) and Node 3 (Variable 3), suggesting that changes in Variable 1 can affect these subsequent variables.
- Acyclic Nature: The graph is acyclic, with no loops, ensuring clear, directional cause-and-effect relationships.
- Quantitative Component: Accompanying Conditional Probability Tables (not shown in the diagram) quantify these relationships, allowing for the computation of the likelihood of each variable's state given the states of its parents.

- Bayesian Inference: This network supports Bayesian inference, updating the probabilities for hypotheses as evidence is incorporated, which is key in scenarios with uncertainty or incomplete information.

Components of a Bayesian Belief Network:

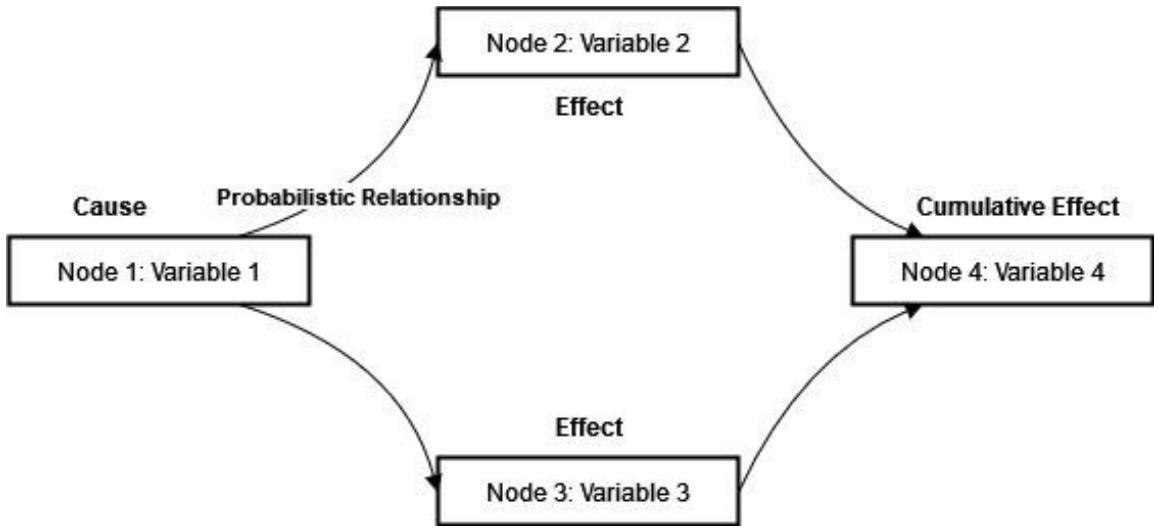

Figure 1: BBN indicating the causal relationship/influence between nodes.

2.2 Application: Pandemic Management

The study conducted in [3] introduces an Exploratory–Descriptive–Explanatory (EDE) framework utilizing evolutionary search algorithms, Bayesian belief networks, and innovative interpretation techniques. This method aims to rapidly identify important chronic risk factors and understand their relationships, enhancing clinical decision-making in the face of novel, complex diseases.

2.2.1 Data

- Data Source: Utilized early US pandemic data from Cerner HealthFacts, a comprehensive EHR platform.
- Dataset Features:
    - Demographics: Age, Gender, Race, Ethnicity.
    - Insurance & Visit: Type, Number of Diagnoses (Num_Dx), History of Diagnoses (Hist_Dx), Visit Type.
    - Chronic Conditions: Over 1000 binary features indicating the presence of chronic conditions.
- Data Preparation: Included only adult patients diagnosed with COVID-19. Performed one-hot encoding for chronic conditions identified using the Chronic Condition Indicator (CCI) list for ICD-10-CM.
- Target Variable: A binary variable indicating patient survival (0 for survived, 1 for deceased)

2.2.2 Exploratory Phase: Feature Selection

- Initial Assumption: The research began with the premise that no specific COVID-19 risk factors were known, treating each chronic condition as a potential risk factor. This approach aimed to tackle the high-dimensional dataset challenge prevalent in bioinformatics.
- Feature Selection through Genetic Algorithm (GA): To manage the dataset's complexity, a GA was employed to identify a subset of features (chronic conditions) that could potentially influence COVID-19 outcomes. In the feature selection process, they utilize Genetic Algorithms (GAs) to optimize the set of features predictive of COVID-19 outcomes. Genetic Algorithms are inspired by evolutionary biology but here, they are applied in a computational context to iteratively improve feature sets.

2.2.3 Descriptive Phase: Understanding Patterns and Inter-Relationships

This phase leveraged descriptive analytics to investigate the relationships and patterns among the selected features and the target variable (survival or death of patients). It involved:

- Analyzing survival and death rates across patients with different chronic conditions, validating the efficacy of the feature selection process.
- Examining comorbid relationships and their impact on COVID-19 outcomes, providing insights into the vulnerability of patients with specific condition pairs.

2.2.4 Explanatory Phase: Probabilistic Prediction Model and Inference Simulator

Utilization in Research:

The research utilizes Bayesian Belief Networks (BBNs) to model probabilistic relationships among risk factors for COVID-19 outcomes. BBNs are graphical models that represent variables as nodes and conditional dependencies as arcs within a Directed Acyclic Graph (DAG). In this study, BBNs are specifically employed to:

- Infer the likelihood of patient outcomes (such as survival or mortality) based on a complex interplay of demographic factors and pre-existing conditions.
- Capture the dynamics between variables, allowing the network to model how changes in one variable affect the probability distribution of another.

Output and Interpretation:

The output of the BBN is a probabilistic framework that predicts outcomes based on the presence or absence of various risk factors. The interpretation of the BBN is two-fold:

- It provides a visual representation of the strength of relationships between risk factors.
- It assists in decision-making by highlighting which factors are crucial for predicting outcomes, thus aiding in the efficient allocation of medical resources.

Mutual Information (MI):

Mutual Information (MI) is a measure derived from information theory, quantifying the information gained about one random variable through another. MI is mathematically defined as the difference in entropy (H) of one variable and the conditional entropy of that variable given another. The formula from the paper is:

$$\mathrm{MI}(\mathrm{Deceased}, \mathrm{Age\_at\_encounter}) = \mathrm{H}(\mathrm{Deceased}) - \mathrm{H}(\,\mathrm{Deceased} \mid \mathrm{Age\_at\_encounter}\,)$$

In general terms for any two variables A and B:

$$\mathrm{MI}(\mathrm{A}, \mathrm{B}) = \mathrm{H}(\mathrm{A}) - \mathrm{H}(\,\mathrm{A} \mid \mathrm{B}\,)$$

Example and Explanation:

An example from the paper is the MI between the 'Age at encounter' and the outcome 'Deceased'. If 'Age at encounter' is known, the MI measures how much this knowledge reduces uncertainty about the patient's outcome. The value is calculated using the probabilities of the patient's age and the outcomes, indicating how much knowing the age contributes to understanding the risk of mortality.

Implications:

MI has significant implications in the study:

- It identifies which variables (risk factors) are most informative for the outcome, which is essential for constructing a dependable BBN model.
- High MI values imply that a variable is a strong predictor of the outcome and should be considered when making clinical decisions.

2.3 Connecting Picture and Interpretation

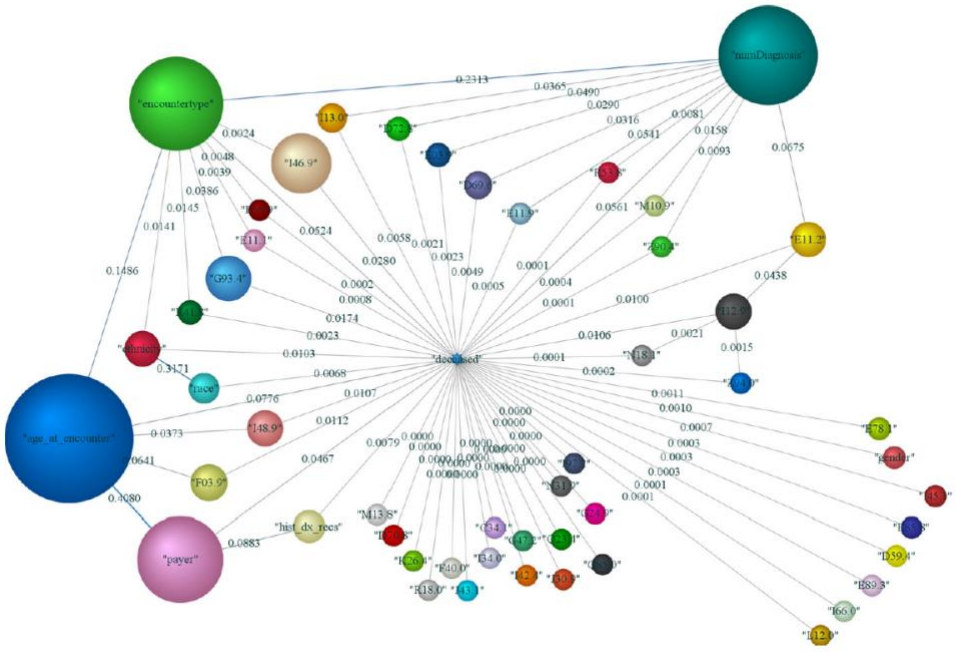

Figure 2: Bayesian Network for risk factors.

The image provided by the authors is a visual representation of the BBN developed in the study. Here is how it connects the concepts of BBN and MI:

- Nodes represent risk factors, and their size is proportional to the MI with the dependent variable (patient outcome).
- Arcs show the MI between nodes, indicating how much one variable tells us about another.

2.4 Results

Summary of Exploratory Analysis:

Category of disease	#Conditions	ICD10 identifier
Renal	6	N18.1; I12.9; N31.9; E11.2; I13.0; Z94.0
Cardiovascular	7	I13.0; I48.9; Q23.4; I46.9; I42.4; I45.1; I34.0
Blood diseases	8	D69.6; I97.2; D72.8; D59.4; D70.8; E11.1; Z86.7; I66.0; Q23.4
Nutritional and metabolism	6	E85.8; E89.3; E55.9; E78.1; E03.9; M10.9
Hypertension	2	I12.9; I13.0
Diabetes	3	E11.1; E11.9; E11.2
Cancer	2	C34.1; R18.0
Lung diseases	2	J43.1; J30.9
Other	11	F03.9; M13.8; F40.0; G93.4; G47.2; K26.4; L12.0; R41.8; R53.8; Q85.0; G24.9

Table 1: Selected chronic risk factors by category.

The authors in [2] delineated chronic conditions implicated in COVID-19 outcomes. Table 1 displays these conditions by category, with their corresponding ICD10 identifiers. The list highlights the diversity of comorbidities, including renal, cardiovascular, and blood diseases, among others, totalling 50 conditions.

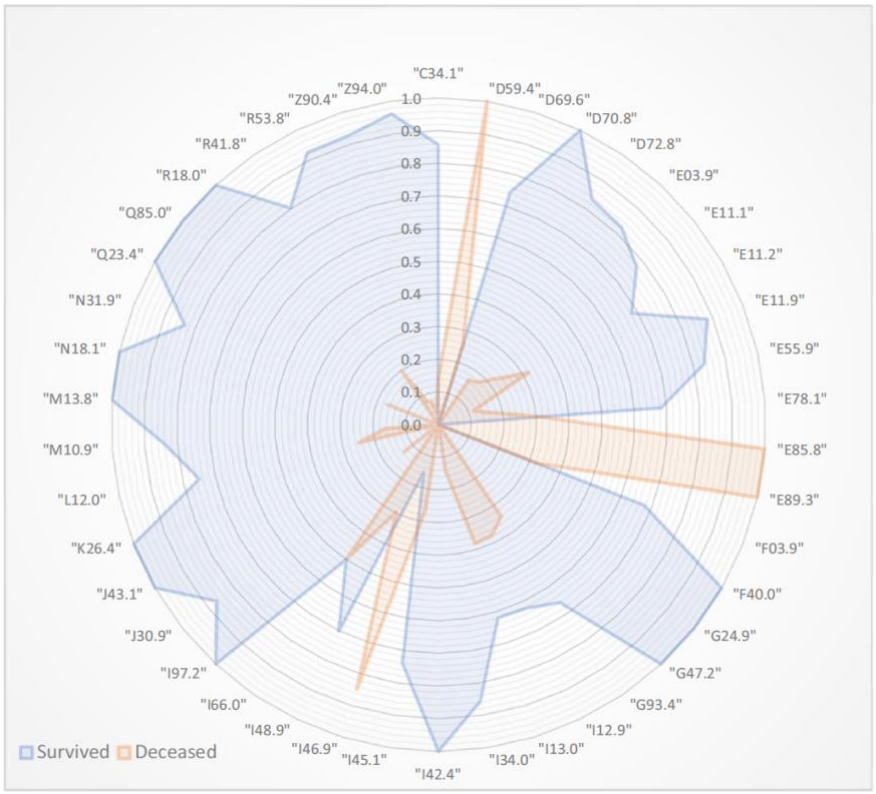

Figure 3: Survival and death rate in COVID-19 patients by their chronic conditions.

They then initiated their exploratory analysis with a feature selection algorithm executed on a high-powered computational setup, which reached convergence after 29 generations. This process resulted in the identification of 50 chronic conditions delineated by category as shown in Figure 2. These conditions, particularly prevalent among senior patient demographics, led to an insightful discussion on age as a potential confounder in the analysis.

Measure	Mean (%)	Min (%)	Max (%)	Standard deviation (%)
K-Fold (10)—Stratified				
Overall precision	90.06	89.27	90.91	0.53
Overall reliability	88.81	87.86	89.57	0.51
Mean ROC index	85.85	82.00	88.30	2.15
Bootstrap (10)				
Overall precision	91.34	91.02	90.63	0.21
Overall reliability	90.38	89.97	90.67	0.93
Mean ROC index	90.78	90.13	91.46	0.39

Table 2: Cross-Validated Performance of Classification of Risk Factors.

Furthering their analytical pursuit, the authors employed Bayesian Networks (BNs) to navigate the complexities inherent in clinical data, especially under conditions of high uncertainty. Incorporating both k-fold stratified cross-validation and bootstrap methods, the authors provided a comprehensive evaluation of the model's credibility. The superior performance of the bootstrap approach, as reported by the authors in Table 2, with a mean ROC index of 90.8 percent compared to 85.8 percent for k-fold, is indicative of the robustness of their probabilistic models.

3. Machine Learning in Drug Discovery, Design and Development

Drug is a very much essential substance in the health care system. Producing a new drug for a disease in the market using traditional methods is very time-consuming and expensive [10]. Therefore CADD, or Computer Aided Drug Design is a discipline that combines computational methods and tools with biology, chemistry, and pharmacology to expedite the drug discovery and development process [10].

Diseases are identified at the molecular level as target molecules. The disease is cured using neutralizing the target molecule interacting with an active drug compound [10]. For various diseases there exist various targets as well as various active drug compounds [10]

The paper mentions that traditionally, most studies have only looked at whether a molecule is active or inactive and that binary classification was most prominently used. However, in this paper, the researchers are interested in a more nuanced classification that includes an additional category called inconclusive [10]. This is a major requirement that is deemed useful in drug discovery as it avoids misclassification and guides it to be further investigated, which is crucial while identifying molecular states.

3.1 Naïve Bayes Classifier vs KNN Algorithm

Paper 1 discusses the performance of the Naïve Bayes Classifier and KNN algorithm on a dataset of compounds that have undergone autofluorescence test (qHTS assay to test for compound autofluorescence at 460 nm (blue) in HEK293 cells), where each record belongs to one of the following classes: “active”, “inactive” or “inconclusive”.

The paper concluded after evaluating several metrics such as Confusion Matrix, Accuracy, Precision, Recall, and F1-score that k-NN is a better classifier than NB for the multi-class drug molecule classification.

A few possible factors that may have influenced KNN being the better one in this case are:

- Drug molecule identification often involves classifying molecules based on complex interactions with biological targets. The non-parametric nature of k-NN allows it to capture these complex relationships, potentially leading to better identification of drug-like molecules compared to Naïve Bayes.
- Naïve Bayes assumes feature independence, which may not hold in many real-world datasets with complex relationships. k-NN's ability to capture non-linear relationships can be advantageous in identifying molecules with specific bioactivities, even in continuous spaces.

3.2 De Novo Drug Design

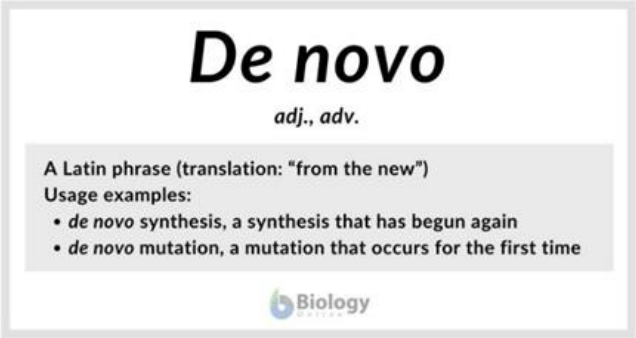

Figure 4: De novo [17]

Over the past decade, significant advancements have been made in machine learning for de novo drug design, particularly in the generative modelling of new chemical structures. However, current generative approaches face a significant challenge: they do not ensure the feasible synthesis of proposed molecular structures, nor do they provide the synthesis routes for the proposed small molecules, thus limiting their practical applicability.

3.2.1 Current Challenges in De Novo Drug Design

- Feasible Synthesis: Current generative models do not ensure that the proposed molecular structures can be feasibly synthesized [12]. This means that the generated structures may not be chemically viable or practical to produce in a laboratory setting.
- Exploitation of Heuristics: Generative models in drug design often rely on heuristics or rules to prioritize the generation of molecules with high accessibility scores [13]. However, these heuristics can be exploited by the generator, leading to the generation of molecules that are still impossible or challenging to produce in practice [12].
- Disjoint Search Pipeline: The use of computer-aided synthesis planning programs creates a disjointed search pipeline. This means that there is a separation between the algorithm used for molecule generation and the algorithm used for synthesizability assessment. This disjointed pipeline does not guarantee that the generative model learns anything about synthesizability.
- Lack of Guarantees: Relying on external synthesis planning programs does not guarantee that the generative model will learn about synthesizability [13]. While these programs can provide valuable guidance, they do not ensure that every molecule proposed by the algorithm can be easily produced.
- Need for Embedded Synthetic Knowledge: There is a need to directly embed synthetic knowledge into de novo drug design. This would allow for the constraint of the search space to synthetically accessible routes and theoretically guarantee that any molecule proposed by the algorithm can be easily produced.

3.3 Reinforcement Learning for Drug Structure Synthesis

In this study, a novel reinforcement learning (RL) setup for de novo drug design, Policy Gradient for Forward Synthesis (PGFS), is proposed. This approach addresses the challenge by incorporating the concept of synthetic accessibility directly into the de novo drug design system.

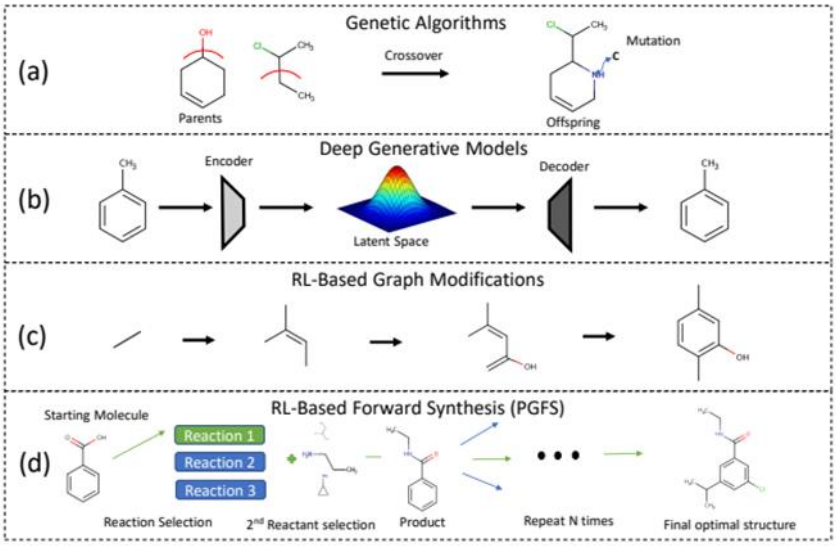

Figure 5: Comparison of Models for Optimal Drug Molecular Structure [12]

Intuitively, the modification or optimization of a molecule can be done in a stepwise fashion, where each step belongs to one of the following three categories: (1) atom addition, (2) bond addition, and (3) bond removal [13]. This suggests that the actions and states in molecular transformations under reinforcement learning

encompass numerous atom and molecule additions and removals, resulting in the generation of new molecular compounds with distinct properties at each step.

The agent learns to select the best set of reactants and reactions to maximize the task-specific desired properties of the product molecule, i.e., where the choice of reactants is considered an action, and a product molecule is a state of the system obtained through a trajectory composed of the chosen chemical reactions. [12]

3.3.1 Pipeline Setup and Methodology for PGFS Model

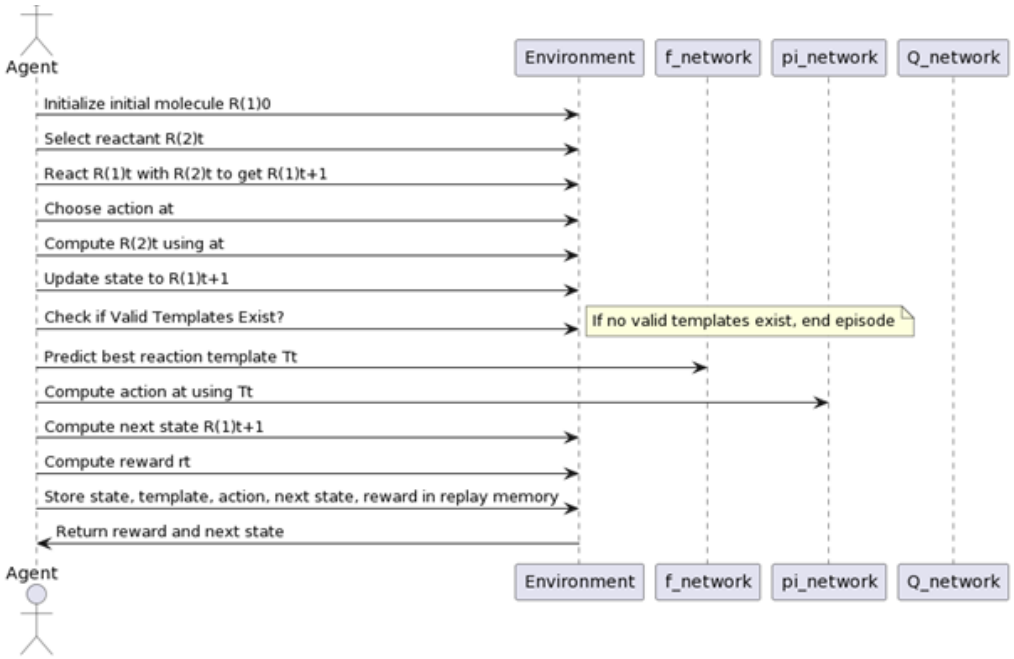

Figure 6: Overview of Policy Gradient for Forward Synthesis (PGFS) model

Initial Molecule Selection:

Initially, at the first time step, the pipeline randomly samples the initial molecule R(1)0 from a list of all commercially available reactants.

At each time step t, the goal of the pipeline is to select a reactant R(2)t to react with the existing molecule R(1)t. Each step also involves choosing the reaction template based on the current state and is determined by the f network, which predicts the best reaction template Tt given the current state (R(1)t).

State and Action:

The reactant R(1)t is considered as the current state st.

- The RL agent chooses an action ‘at’, which is used to compute the reactant R(2)t.
- The product R(1)t+1, considered as the next state st+1, is determined by the environment based on the two reactants R(1)t and R(2)t.
- A novel Markov Decision Process (MDP) involving a continuous action space is formulated to handle the challenge of many possible reactants.

Actor-Critic Framework:

The agent comprises three learnable networks: f, ⇡, and Q. The actor module ⇧ includes f and ⇡ networks, while the critic is the Q network estimating the Q-value of the state-action pair.

Reaction Template Selection:

To overcome the limitation of large discrete action space where there are over a hundred thousand possible second reactants, an intermediate action is introduced by choosing a reaction template encoded in the SMARTS language.

The f network performs the choice of reaction template and constrains the space of possible R(2)s to those containing a particular substructure.

To ensure effective gradient propagation through f network the template T is multiplied with the template mask Tmask, after which Gumbel SoftMax is applied.

$$T = T \odot T_{mask}$$

$$T = GumbelSoftmax(T, \tau)$$

k-NN for Action Selection:

After choosing a reaction template, the action space is constrained to those reactants that contain a specific substructure required by the reaction template. However, even with these constraints, tens of thousands of reactants can still be at each step. To handle this, the researchers use the k-nearest neighbours (k-NN) technique to select k reactants from the set of all possible reactants R(2) that are closest to the action a.

Actor Module:

- At each time step t, the actor module takes the current state st (R(1)t) as input and produces an action 'at'. The action 'at' is a tensor, defined in the feature representation space of all initial reactants R(2).
- The f network predicts the best reaction template Tt given the current state st.
- Using Tt and st, the ↟ network computes the action 'at'.

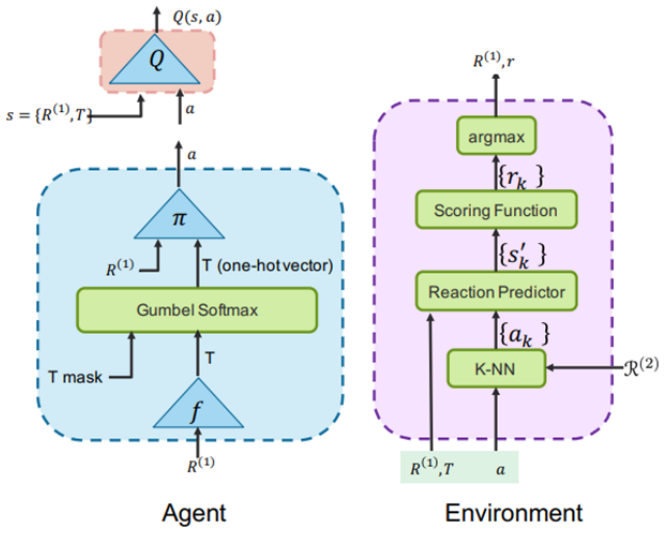

Figure 7: PGFS Environment and Agent [12]

Environment Response:

- The environment takes st, Tt, and 'at' as inputs and computes the reward rt, next state st+1, and a Boolean to determine if the episode has ended.
- It selects k reactants from the set R(2) closest to the action 'at' using the k nearest neighbours technique.
- The rewards associated with the products are computed using a scoring function, and the reward and product corresponding to the maximum reward are returned.
- The state st, best template Tt, action 'at', next state st+1, and reward rt are stored in the replay memory buffer.

Episode Termination:

The episode terminates when the maximum number of reaction steps is reached or when the next state has no valid templates.

3.4 Results and Analysis

The performance of PGFS is evaluated across various scoring functions that measure drug-likeness, biological activity, and synthesizability. This section presents a comparative analysis based on maximum achieved values with these scoring functions. In the recent advancements made to the Policy Gradient for Forward Synthesis (PGFS) model, evaluating its performance against established benchmarks in the domain of drug design provided insightful results.

The model was tasked with navigating the complex chemical space to identify small molecules that could be synthesized and potentially serve as drug candidates. Scoring functions reflecting drug-likeness (QED) and synthetic feasibility (penalized clogP), alongside biological activity against HIV-related targets (RT, INT, CCR5), served as metrics for evaluation.

A comparison was made between the PGFS model's performance, and the baseline set by the ENAMINEBB database, as well as other state-of-the-art models such as Random Search (RS), Graph Convolutional Policy Network (GCPN), Junction Tree Variational Autoencoder (JT-VAE), and Multi-Objective Optimization (MSO). The ability of each method to produce compounds with optimal scores in each metric was assessed.

Method	QED	clogP	RT	INT	CCR5
ENAMINEBB	**0.948**	5.51	7.49	6.71	8.63
RS	**0.948**	8.86	7.65	7.25	8.79 (8.86)
GCPN	**0.948**	7.98	7.42(7.45)	6.45	8.20(8.62)
JT-VAE	0.925	5.30	7.58	7.25	8.15 (8.23)
MSO	**0.948**	26.10	7.76	7.28	8.68 (8.77)
PGFS	**0.948**	**27.22**	**7.89**	**7.55**	**9.05**

Table 3: Performance comparison of the maximum achieved value with different scoring functions. [12]

Results indicated that the PGFS model consistently outperformed the random baseline and displayed competitive or superior results in generating molecules with high scores in the QED and penalized clogP metrics. Specifically, PGFS showcased a significant enrichment of high-scoring compounds, denoting a substantial shift in the score distribution towards higher values and affirming the model's effective learning and predictive capabilities.

In the realm of HIV-related targets, PGFS demonstrated the ability to synthesize molecules with potentially higher efficacy, as evidenced by its performance on the CCR5 reward function. This implies that molecules generated by PGFS were more likely to inhibit HIV-related biological processes effectively compared to those produced by other methods.

	NO AD			AD		
Scoring	RT	INT	CCR5	RT	INT	CCR5
ENAMINEBB	6.87 ± 0.11	6.32 ± 0.12	7.10 ± 0.27	6.87 ± 0.11	6.32 ± 0.12	6.89 ± 0.32
RS	7.39 ± 0.10	6.87 ± 0.13	8.65 ± 0.06	7.31 ± 0.11	6.87 ± 0.13	8.56 ± 0.08
GCPN	7.07 ± 0.10	6.18 ± 0.09	7.99 ± 0.12	6.90 ± 0.13	6.16 ± 0.09	$6.95^{*} \pm 0.05$
JT-VAE	7.20 ± 0.12	6.75 ± 0.14	7.60 ± 0.16	7.20 ± 0.12	6.75 ± 0.14	7.44 ± 0.17
MSO	7.46 ± 0.12	6.85 ± 0.10	8.23 ± 0.24	7.36 ± 0.15	6.84 ± 0.10	$7.92^{*} \pm 0.61$
PGFS	$\mathbf{7.81} \pm 0.03$	$\mathbf{7.16} \pm 0.09$	$\mathbf{8.96} \pm 0.04$	$\mathbf{7.63} \pm 0.09$	$\mathbf{7.15} \pm 0.08$	$\mathbf{8.93} \pm 0.05$

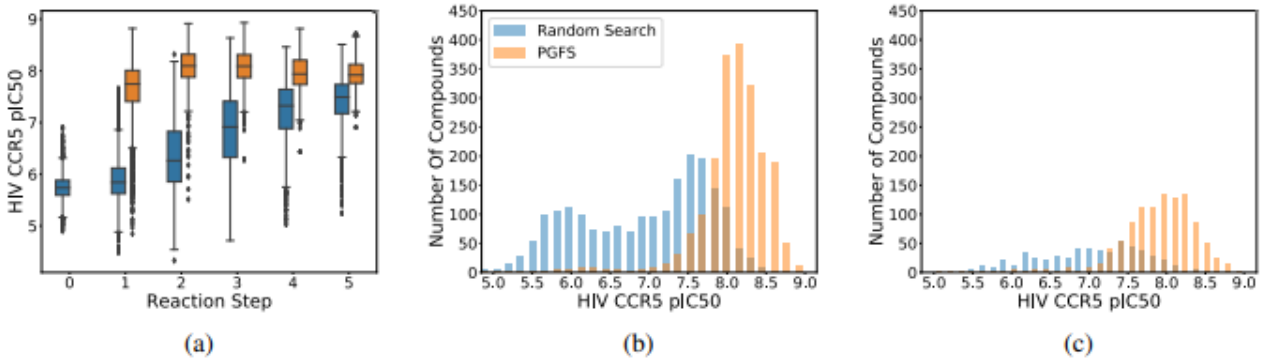

Figure 8: Performance comparison between Random Search and PGFS [12]

The quantitative benchmarking against peer models reaffirmed the robustness of PGFS in maximizing the scores for the defined tasks. Notably, the unrestricted nature of PGFS regarding molecular size and its reflection on the penalized clogP score accentuates the model's adeptness at exploring chemical space without stringent constraints.

In conclusion, the analysis reveals PGFS as a potent tool in the de novo drug design landscape, with promising implications for the future of medicinal chemistry and the discovery of new therapeutic agents. The findings encourage further exploration and refinement of the model to harness its full potential in the automated design of novel drug candidates.

4. Machine Learning for Enhanced Diagnosis through Genetic Profiling

4.1 Machine Learning in Genetic Healthcare

The increasing role of machine learning (ML) in healthcare, particularly in genetic data analysis, is transformative. ML's capacity to analyze complex genetic data is unlocking new possibilities for disease diagnosis, enhancing treatment options, and improving patient outcomes [5]. Through the identification of genetic markers and patterns, ML facilitates personalized medicine approaches, tailoring treatments to individual genetic profiles.

Support Vector Machine (SVM) is a state-of-the-art classification technique using multiple features to establish a two-group classification. It works by finding the optimal hyperplane that best separates different classes in a high-dimensional feature space. SVM is widely used due to its effectiveness in handling complex datasets, ability to handle high-dimensional data, and flexibility in using different kernel functions for nonlinear classification.

4.2 Genetic Markers in Alzheimer's Disease Progression

Alzheimer's is a disorder of the brain resulting in memory loss, and cognitive and intellectual impairments able to affect social activities and decision-making. Identifying genetic variants associated with complex diseases is considered one of the most important studies about the human genome. Genetic association studies aim to identify SNPs that are most associated with complex and common diseases to improve early diagnosis and treatment of these diseases [7]. Single Nucleotide Polymorphisms (SNPs) are the most common type of human genetic variation.

In this study, the SVM classifier was utilized to differentiate patients with Alzheimer's disease (AD) from control subjects, aiming to identify genetic biomarkers associated with AD [7]. The training of the data was conducted using the Sequential Minimal Optimization (SMO) algorithm, chosen for its proficiency in solving optimization problems and generating precise outcomes efficiently. It breaks down the optimization problem into smaller subproblems, making it suitable for handling large datasets and complex classification tasks.

4.3 Methodological Approach to Genetic Profiling in Alzheimer's

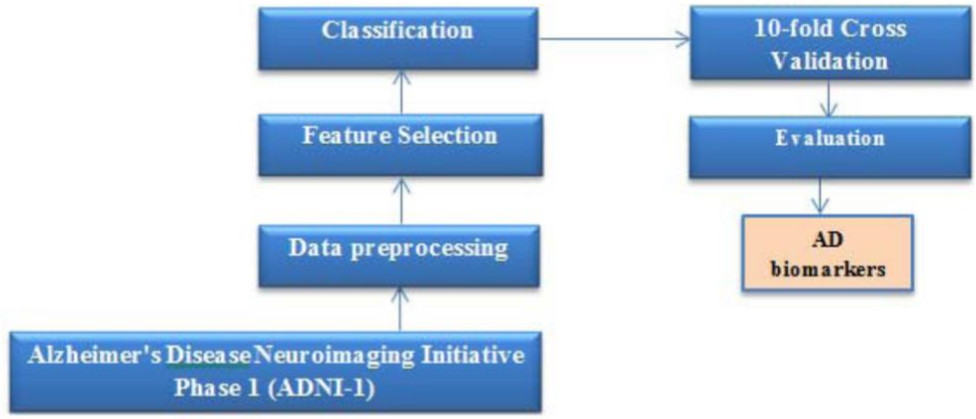

Figure 9: Workflow [7]

The workflow described in Figure 9 starts with the collection of genetic and phenotype data from the ADNI-1 dataset. Apolipoprotein E (APOE) is a gene that has been confirmed as the most important risk factor for AD listed on the AlzGene database. APOE genotyping refers to genetic testing for determining the variations of the

APOE gene, which have been linked to the risk of developing AD. Individuals in the dataset are categorized as cognitively normal (CN), having mild cognitive impairment (MCI), or Alzheimer's disease (AD). After applying Quality Control (QC) to the genetic data, a set of SNPs is selected according to the top ten ranked genes listed on the AlzGene database. The total number of selected SNPs is 115 SNPs. Feature selection methods are applied to identify the most significant SNPs. SMO is used to train SVM classifiers with different kernels to discriminate between CN and AD individuals based on the selected SNPs. The performance of these classifiers is then validated using 10-fold cross-validation, and their effectiveness is measured in terms of accuracy, recall, and precision to identify the most significant SNPs associated with AD.

4.3.1 Data Collection

The data from the Alzheimer's Disease Neuroimaging Initiative Phase 1 (ADNI-I), focusing on Single Nucleotide Polymorphisms (SNPs) associated with AD was selected. ADNI-1 is a large-scale project that collects and analyses medical data to understand the progression of Alzheimer's Disease better. This initiative provides a dataset that includes genetic information, which can be used to identify biomarkers for AD. ADNI-l data is gathered from 757 total individuals including 214 CN, 366 MCI and 177 AD recruited from more than 50 sites across the United States and Canada [6].

4.3.2 Data Integrity and Genetic Insights

The study focuses on the APOE gene, which is a significant genetic marker for Alzheimer's Disease (AD). It analyses two crucial SNPs, rs429358 and rs7412, to determine the APOE genotype of individuals, providing insights into their AD risk levels. These specific SNPs are chosen for their established link to AD. The research also examines additional SNPs to further investigate genetic contributions to the disease. Participants were categorized as either cognitively normal (CN) or afflicted with Alzheimer's (AD), designated numerically as 1 or 2, respectively. To maintain the dataset's integrity for robust genetic analysis, participants with incongruent gender data and SNPs with block inheritance were excluded, while closely related individuals and data entries with substantial missing information were also omitted to ensure data quality and independence. After cleaning, the dataset was refined to include 751 subjects—211 CN, 365 with mild cognitive impairment (MCI), and 175 AD and 566,016 SNPs to analyze for links to AD.

4.3.3 Optimizing SNP Selection for Alzheimer's Prediction

The goal of feature selection is to choose a subset of SNPs that have the strongest relationship with AD while avoiding redundancy (where SNPs provide the same information), which helps in building a more accurate and efficient classifier. Researchers start with a list of candidate genes from the AlzGene database [7], and they extract 115 SNPs believed to be relevant for the disease.

Correlation-based Feature Selection (CFS) is used to search for the SNPs that are highly correlated with the class and have low inter-correlation with each other by assigning high scores [8]. The Greedy Stepwise search algorithm [9] searches the entire list of features, considering their predictive ability to sort them. As a result of that step, a ranked list of the SNPs is obtained. After that, the best possible SNPs are selected, including 21 SNPs. The Chi-Squared Attribute Evaluator with a ranker search method is applied for ranking features. Chi-squared attribute evaluation evaluates the worth of an attribute by computing the chi-squared statistics for the class. The ranker search method is used according to their relevance. A threshold is set to eliminate some unwanted features. The final output is a list of 21 SNPs selected by both CFS and Chi-Squared methods, the corresponding gene, chromosome (Chr), and potential pathways these genes are involved in, such as lipid metabolism, immune response, and endocytic pathways. All these pathways are believed to have roles in AD.

4.3.4 Use of SMO Algorithm for the training of SVM

SVMs work by finding the best-separating hyperplane that divides data points from different classes in the feature space. Training an SVM requires solving quadratic programming problems, which involve finding an optimal point under certain constraints. These are typically complex and computationally intensive. SMO algorithm simplifies the training of SVMs. Instead of solving a large QP problem all at once, SMO breaks it

down into smaller QP sub-problems that can be solved more efficiently. This makes the process faster because it avoids complex numerical QP optimization steps.

Kernel functions are used for finding data relations in the dataset. Kernel methods enable data to operate in higher dimensional space by detecting the kernel function that is appropriate for higher classification accuracy. By applying kernel functions, researchers can find relationships and patterns in genetic data that might not be apparent in the original feature space, aiding in the discovery of genetic variants linked to diseases. In this study, four kernel functions were used: linear, polynomial, RBF, and PUK.

4.3.5 Validating Predictive Models with Cross-Validation

Cross-validation is a model validation technique used to assess how the results of a statistical analysis will generalize to an independent data set. 10-fold cross-validation is applied to test the performance of the Support Vector Machine (SVM) classifier to identify Alzheimer's Disease (AD) biomarkers.

4.4 Evaluation Metrics

The metrics used are accuracy (ACC), recall (REC), and precision (PREC).

- Accuracy: The percentage of all correct predictions.
- Recall: The proportion of actual positive cases (AD cases) correctly identified.
- Precision: The proportion of correct identifications (how many cases identified as AD by the model were indeed AD).

4.5 Results

Table 4 summarizes the best selected and ranked SNPs using CFS with Greedy-stepwise search algorithm and Chi-Squared Attribute Evaluator with Ranker algorithm, their gene name, chromosome number (chr), and potential pathways. These genes are the top candidate genes most associated with AD that are listed on the AlzGene database. SNPs that are highly associated with AD are identified in the APOE, ABCA7, BINI, CD2AP, CD33, CLU, CRI, MS4A6A and PICALM genes.

Selected SNPs		Gene	Chr	Potential pathways
CFS	Chi-Squared			
rs429358	**rs429358**	APOE	19	Lipid metabolism Cholesterol metabolism
rs2074447 **rs2074451**	**rs2074451**	ABCA7	19	Cholesterol metabolism Lipid metabolism Immune system Inflammatory response
rs1595816	**rs1595816**	BIN1	2	Endocytic pathways
rs13207896 **rs9296559** rs9395267	**rs13207896** **rs9296559**	CD2AP	6	Endocytic pathways Inflammatory response Immune system
rs2072561 **rs12982353** **rs2007332**	**rs2072561** **rs12982353** **rs2007332** rs4803892 rs2041992	CD33	19	Immune system Inflammatory response
rs512941	**rs512941**	CLU	8	Immune system Cholesterol metabolism Lipid metabolism
rs2660635 rs17014396 rs12119464	**rs2660635**	CR1	1	Immune system Inflammatory response
rs11230442	**rs11230442**	MS4A6A	11	Immune system Inflammatory response
rs1945043 **rs7950878** rs1945146 **rs7940587** **rs659301** **rs661954**	**rs1945043** **rs7950878** **rs7940587** **rs659301** **rs661954** rs7103017 rs4944603 rs722158	PICALM	11	Endocytic pathways

Table 4: Top 21 selected SNPs [7]

Table 5 shows the SMO classifier's classification accuracy, precision, and recall. The Chi-Squared feature selection algorithm achieves higher accuracy than CFS with SMO trained using Linear, Quadratic Polynomial and Cubic Polynomial kernel models. Both feature selection methods show similar results with SMO trained using the PUK kernel model. The highest accuracy (76.70%) is observed when using the RBF kernel model, irrespective of the feature selection method used, indicating its effectiveness in this classification task for Alzheimer's disease.

SMO kernel models	CFS with Greedy-stepwise search algorithm			Chi Squared with Ranker algorithm		
	ACC %	REC	PREC	ACC %	REC	PREC
Linear	75.23	0.75	0.60	75.63	0.76	0.64
Quadratic Polynomial	61.65	0.62	0.64	65.65	0.66	0.67
Cubic Polynomial	63.38	0.63	0.64	68.44	0.68	0.67
RBF	76.70	0.77	0.59	76.70	0.77	0.59
PUK	76.56	0.77	0.59	76.56	0.77	0.59

Table 5: Classification Results (ACC: Accuracy; REC: Recall; PREC: Precision) [7]

5. Future Outlook

The progress made in Machine Learning (ML) and Artificial Intelligence (AI) has made a significant impact on the field of healthcare, specifically in the areas of pandemic management, genetic profiling, and drug discovery. There is immense potential for further advancements in these domains, presenting numerous opportunities for innovation and enhancement.

In the realm of pandemic management, the rapid development and widespread implementation of mRNA vaccines have showcased the effectiveness of ML and AI in controlling outbreaks such as the COVID-19 pandemic. However, there are still challenges to address, including the long-term safety monitoring of these novel vaccines and the efficient distribution of vaccines. Future efforts should focus on improving ML algorithms to detect any potential long-term effects and adverse events, thereby ensuring that public health policies can be adjusted based on new data. Moreover, AI can play a critical role in optimizing the logistics of vaccine distribution, ensuring that resources are allocated effectively to safeguard populations against future health threats.

When it comes to developing new drugs, one approach called PGFS has shown it can effectively search the space of possible drug molecules. Going forward, they may want to incorporate other reinforcement learning methods too, like Soft Actor-Critic, to better explore the chemical options. Additionally, developing a second policy gradient could lead the system to pick the best chemical transformations and stop when it thinks it's found the best possible drug candidate. Together, these enhancements may significantly boost how efficiently and effectively we can design new drugs from scratch, potentially finding treatments with ideal properties.

When it comes to genetic profiling, Support Vector Machines (SVMs) have exhibited promise in identifying Single Nucleotide Polymorphisms (SNPs) that are linked to complex diseases. SVMs utilize ML techniques to analyze vast datasets and identify relevant SNPs that could provide valuable insights into the genetic basis of various illnesses. By leveraging ML and AI in genetic profiling, researchers can uncover crucial information that may lead to improved disease prevention, diagnosis, and treatment strategies.

ML and AI have revolutionized the field of drug discovery by enabling researchers to analyze vast amounts of data and identify potential drug candidates more efficiently. By employing algorithms that can recognize patterns and predict molecular interactions, ML and AI accelerate the identification of promising compounds for further development. This not only expedites the drug discovery process but also increases the likelihood of finding effective treatments for various diseases.

6. Conclusion

The advancements in ML and AI have had a profound impact on healthcare, particularly in pandemic management, genetic profiling, and drug discovery. The future holds immense potential for further progress in these areas, offering numerous opportunities for innovation and improvement. By harnessing the power of ML and AI, we can continue to transform healthcare and improve the well-being of individuals worldwide.

Researchers are working on improving algorithms to uncover hidden patterns in genetic data. This could help discover genetic variants linked to diseases. In turn, that would help make diagnoses more accurate and treatments more tailored to everyone.

In summary, blending machine learning and AI with healthcare seems promising for dealing with challenges like pandemics, analyzing genes, and discovering drugs. By taking advantage of these technologies, the future could bring healthcare that's more precise for each person, based on their own traits and genes. Ultimately, that level of customization may translate to better outcomes for patients and improved public health overall.

7. References

[1] Sudakov VA, Titov YP. Pandemic Forecasting by Machine Learning in a Decision Support Problem. Math Models Comput Simul. 2023;15(3):520–8. doi: 10.1134/S2070048223030171. Epub 2023 May 17. PMCID: PMC10191073.

[2] Topuz K, Davazdahemami B, Delen D. A Bayesian belief network-based analytics methodology for early-stage risk detection of novel diseases. Ann Oper Res. 2023 May 17:1-25. doi: 10.1007/s10479-023-05377-4. Epub ahead of print. PMID: 37361089; PMCID: PMC10189691.

[3] Chapter 7, Barbrook-Johnson, P., Penn, A.S. (2022). Bayesian Belief Networks. In: Systems Mapping. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-01919-7_7

[4] Sarmiento Varón L, González-Puelma J, Medina-Ortiz D, et al. The role of machine learning in health policies during the COVID-19 pandemic and in long COVID management. Front Public Health. 2023;11. doi: https://doi.org/10.3389/fpubh.2023.1140353

[5] L. Alajramy, A. Taweel, R. Jarrar, E. Lamine and I. Megdiche, "Automated Learning Approach for Genetic Diseases," in 2022 IEEE/ACS 19th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2022 pp. 1-6.

[6] L. Bertram, C. M. Lill, and R. E. Tanzi, "The genetics of Alzheimer disease: back to the future", Neuron, vol. 68, No.2, pp. 270-281, 2010.

[7] M. Mostafa Abd El Hamid, Y. M. K. Omar and M. S. Mabrouk, "Identifying genetic biomarkers associated to Alzheimer's disease using Support Vector Machine," 2016 8th Cairo International Biomedical Engineering Conference (CIBEC), Cairo, Egypt, 2016, pp. 5-9, doi: 10.1109/CIBEC.2016.7836087.

[8] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning," Hamilton, New Zealand, 1998.

[9] X. Geng, T. Y Liu, T. Qin, H. Li, Qin, "Feature selection for ranking" Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 407-414, 2007.

[10] L. Mandal and N. D. Jana, "A Comparative Study of Naive Bayes and k-NN Algorithm for Multi-class Drug Molecule Classification," 2019 IEEE 16th India Council International Conference (INDICON), Rajkot, India, 2019, pp. 1-4, doi: 10.1109/INDICON47234.2019.9029095.keywords: {Drugs; Compounds; Measurement; Diseases; Bayes methods; Proteins; Training; CADD; Drug Design; k-NN and Naive Bayesian}, A Comparative Study of Naive Bayes and k-NN Algorithm for Multi-class Drug Molecule Classification | IEEE Conference Publication | IEEE Xplore

[12] Gottipati, S.K., Sattarov, B., Niu, S., Pathak, Y., Wei, H., Liu, S., Liu, S., Blackburn, S., Thomas, K., Coley, C., Tang, J., Chandar, S. & Bengio, Y.. (2020). Learning to Navigate the Synthetically Accessible Chemical Space Using Reinforcement Learning. Proceedings of the 37th International Conference on Machine Learning, in Proceedings of Machine Learning Research 119:3668-3679 Available from https://proceedings.mlr.press/v119/gottipati20a.html.

[13] Zhou, Z., Kearnes, S., Li, L. et al. Optimization of Molecules via Deep Reinforcement Learning. Sci Rep 9, 10752 (2019). https://doi.org/10.1038/s41598-019-47148-x

[14] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2019). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971v6. Submitted on 9 Sep 2015 (v1), last revised 5 Jul 2019 (this version, v6).[1509.02971] Continuous control with deep reinforcement learning (arxiv.org)

[15] A. A. Jumaily, M. Mukaidaisi, A. Vu, A. Tchagang and Y. Li, "Exploring Multi-Objective Deep Reinforcement Learning Methods for Drug Design," 2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Ottawa, ON, Canada, 2022, pp. 1-8, doi: 10.1109/CIBCB55180.2022.9863052. keywords: {Drugs; Training; Proteins; Scalability; Design methodology; Reinforcement learning; Lead; reinforcement learning; deep reinforcement learning; multi-objective optimization; drug design; Deep FMPO},Exploring Multi-Objective Deep Reinforcement Learning Methods for Drug Design | IEEE Conference Publication | IEEE Xplore

[16] Dulac-Arnold, G., Evans, R., Sunehag, P., and Coppin, B. Reinforcement learning in large discrete action spaces. ArXiv, abs/1512.07679, 2015. [1512.07679] Deep Reinforcement Learning in Large Discrete Action Spaces (arxiv.org)

[17] De novo - Definition and Examples | Biology Online