Machine Learning with Applications 17 (2024) 100576

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects

George Obaido [a,*], Ibomoiye Domor Mienye [b], Oluwaseun F. Egbelowo [c], Ikiomoye Douglas Emmanuel [d], Adeola Ogunleye [b], Blessing Ogbuokiri [e], Pere Mienye [f], Kehinde Aruleba [g]

[a] Center for Human-Compatible Artificial Intelligence (CHAI), Berkeley Institute for Data Science (BIDS), University of California, Berkeley, CA, 94720, USA
[b] Institute of Intelligent Systems, University of Johannesburg, Johannesburg, 2006, South Africa
[c] Department of Integrative Biology, The University of Texas at Austin, Austin, TX, 78712, USA
[d] School of Science, Engineering and Environment, University of Salford, Salford, United Kingdom
[e] Department of Computer Science, Brock University, Niagara Region, St. Catharines, ON, L2S 3A1, Canada
[f] Health Plus, Lekki, Lagos, Nigeria
[g] School of Computing and Mathematical Sciences, University of Leicester, Leicester, LE1 7RH, United Kingdom

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Machine learning
Neural network
Supervised learning

ABSTRACT

Drug discovery and development is a time-consuming process that involves identifying, designing, and testing new drugs to address critical medical needs. In recent years, machine learning (ML) has played a vital role in technological advancements and has shown promising results in various drug discovery and development stages. ML can be categorized into supervised, unsupervised, semi-supervised, and reinforcement learning. Supervised learning is the most used category, helping organizations solve several real-world problems. This study presents a comprehensive survey of supervised learning algorithms in drug design and development, focusing on their learning process and succinct mathematical formulations, which are lacking in the literature. Additionally, the study discusses widely encountered challenges in applying supervised learning for drug discovery and potential solutions. This study will be beneficial to researchers and practitioners in the pharmaceutical industry as it provides a simplified yet comprehensive review of the main concepts, algorithms, challenges, and prospects in supervised learning.

1. Introduction

Drug discovery and development is a time-consuming process that involves identifying, designing, and testing new drugs to address critical medical needs (Aly & Alotaibi, 2023; Athreya et al., 2019; Ekins et al., 2019; Sarkar et al., 2023; Vamathevan et al., 2019). Once a promising compound is identified, it undergoes rigorous testing in preclinical and clinical trials to assess its safety, efficacy, and potential side effects (Jantan, Ahmad, & Bukhari, 2015; Koivisto, Belvisi, Gaudet, & Szallasi, 2022; Zhou et al., 2016). This process can take years, involving collaboration between scientists, physicians, regulatory agencies, and pharmaceutical companies. Despite these challenges, successful drug discovery can lead to groundbreaking treatments that improve patients' lives and advance medical science.

In recent years, technological advancements have revolutionized various aspects of the drug discovery process, streamlining and accelerating certain stages (Mak, Wong, & Pichika, 2023; Rubin, Tummala, Both, Wang, & Delaney, 2006; Selekman et al., 2017). High-throughput screening techniques allow researchers to rapidly test thousands of compounds for potential therapeutic effects, significantly speeding up the initial identification phase. Additionally, computational methods, such as artificial intelligence are increasingly utilized to predict the properties and behavior of drug candidates, reducing the reliance on traditional trial-and-error approaches (Fu et al., 2024; Janakiraman, Khanna, & Ramkanth, 2023; Marchetti, Moroni, Pandini, & Colombo, 2021; Tayyebi et al., 2023; Wallach, Dzamba, & Heifets, 2015; Zhang & Liu, 2019).

[*] Corresponding author.
E-mail addresses: gobaido@berkeley.edu (G. Obaido), ibomoiyem@uj.ac.za (I.D. Mienye), oluwaseun@aims.edu.gh (O.F. Egbelowo), i.d.emmanuel@edu.salford.ac.uk (I.D. Emmanuel), 216091856@student.uj.ac.za (A. Ogunleye), bogbuokiri@brocku.ca (B. Ogbuokiri), pere.mienye@healthplusnigeria.com (P. Mienye), ka388@leicester.ac.uk (K. Aruleba).

https://doi.org/10.1016/j.mlwa.2024.100576
Received 12 June 2024; Received in revised form 4 July 2024; Accepted 13 July 2024
Available online 24 July 2024

Machine learning (ML), a subset of artificial intelligence (AI), gives systems the ability to learn from data without being explicitly programmed (Rustam et al., 2020). It has played a significant role in recent innovations in various fields. For example, ML algorithms have been utilized to diagnose diseases and predict patient outcomes based on their medical history and lifestyle factors (Aruleba et al., 2020; Cheong, Au Yeung, Quon, Concepcion, & Kong, 2021; Mienye, Sun, & Wang, 2019; Obaido et al., 2024; Obaido, Ogbuokiri, Mienye, & Kasongo, 2022). In finance, ML has been used to develop models for fraud detection, credit risk prediction, credit scorecards, and decision engines (Mienye & Sun, 2023). Furthermore, ML has been used in retail to build models that predict the product a customer is likely to purchase. ML is revolutionizing many industries by simplifying and solving complex problems. The integration of ML and AI in drug discovery has sparked a transformative shift in the pharmaceutical industry, leading to enhanced efficiency and efficacy in the development of novel therapeutics (Abeel, Helleputte, Van de Peer, Dupont, & Saeys, 2010; Dipnall et al., 2016; Ghosh et al., 2022; Gutiérrez-Gómez, Vohryzek, Chiêm, Baumann, Conus, Do Cuenod, Hagmann, & Delvenne, 2020; Hajjo, Sabbah, Bardaweel, & Tropsha, 2021; Rehman, Zhuang, Muhamed Ali, Ibrahim, & Li, 2019; Salvatore et al., 2015; Xie et al., 2021; Zhang, Jonassen, & Goksøyr, 2021; Zhang & Liu, 2019). ML algorithms have been employed across various drug discovery subdomains such as genomics, proteomics, and transcriptomics, uncovering critical molecular pathways and biomarkers associated with different diseases. This has enabled the prioritization and validation of viable drug targets.

Meanwhile, ML algorithms are usually grouped into four main types, including supervised, unsupervised, semi-supervised, and reinforcement learning (Dalal, 2020; Haldorai, Ramu, & Suriya, 2020). Supervised learning (SL) algorithms are employed for problems in which the training data contains labeled samples (Fabris, Magalhães, & Freitas, 2017). In contrast, unsupervised learning involves building models using unlabeled data, and the algorithm is expected to detect patterns and relationships in the data without any pre-defined labels. Semi-supervised learning combines supervised and unsupervised learning, where a few labeled data are used together with a larger number of unlabeled data (Yang, Song, King, & Xu, 2021). Furthermore, reinforcement learning (RL) methods enable an intelligent agent to communicate with the environment and learn via trial and error with feedback from its actions (Gronauer & Diepold, 2022). The feedback can be positive or negative, representing a reward or punishment to maximize the reward function. RL-based models learn from their mistakes, offering AI systems that closely mimic human intelligence.

SL algorithms have wider applications and are more popular (Huang, Chen, Lin, Ke, & Tsai, 2017; Sindhu Meena & Suriya, 2020; Uddin, Khan, Hossain, & Moni, 2019). The most commonly used SL algorithms include logistic regression (LR), decision trees (DT), support vector machines (SVMs), random forest (RF) and neural networks (Esteva et al., 2017; Obaido et al., 2024; Sindhu Meena & Suriya, 2020; Uddin et al., 2019). Each of these algorithms has its strengths and limitations, and the choice of algorithm mainly relies on the specific problem. This survey provides an overview of the main SL algorithms and the challenges and prospects for future research. The survey is important and timely because of the increasing use of ML in various applications and the need for a comprehensive understanding of its capabilities and limitations, together with their mathematical formulations. By understanding the strengths and limitations of the different algorithms, researchers and practitioners can make informed decisions about which approach to use for a given problem.

Fig. 1 describes the different application phases of supervised ML to drug discovery and development. Each phase contributes to a more integrated and targeted approach to drug design and personalized medicine, ensuring that treatments are not only effective but also safer and more tailored to individual needs. Furthermore, despite the widespread use of ML algorithms, researchers and practitioners encounter several challenges to drug discovery and development. This survey addresses some of these challenges, such as overfitting, data imbalance, bias and fairness, and interpretability of models. In particular, by highlighting and discussing these challenges, the survey can guide future research efforts to address these issues and improve the performance and interpretability of SL models in the drug design field.

The rest of the paper is structured as follows: Section 2 reviews related works in recent literature, while Section 3 presents an overview of ML. Section 4 provides a detailed summary of SL algorithms, and Section 5 provides several performance metrics applied for SL, especially for classification and regression tasks. Section 6 describes their application to drug design. Section 7 highlights notable challenges in SL. Section 8 discusses the study's findings and suggests future research directions, while Section 9 concludes the study.

2. Related works

Several researchers have reviewed SL and provided valuable insights. For instance, Singh, Thakur, and Sharma (2016) conducted a comprehensive review that grouped SL algorithms based on their ability to categorize data from prior information. This study emphasized the effectiveness of these algorithms, considering factors such as speed, accuracy, complexity, and the risk of overfitting. Similarly, Osisanwo et al. (2017) presented a review of several SL algorithms, including LR, DT, SVM, and several linear classifiers, and evaluated their performance on both small and large datasets.

Choudhary and Gianey (2017) provided a comprehensive comparison of various SL algorithms, including regression algorithms (linear regression, SVM regressor, DT regressor, and LASSO) and classification algorithms (LR, Naïve Bayes (NB), k-Nearest Neighbors (k-NN), and DT). Their work emphasized that each algorithm is unique, tailored to specific applications, and there is no one-size-fits-all ‘‘powerful’’ algorithm; and that the choice of an algorithm should be contingent on the nature of the task and the available data. Furthermore, Nasteski (2017) reviewed four SL algorithms, namely DT, Linear regression, NB, and LR, and noted that this ML method has remained dominant due to their clearer criteria for model optimization.

Some researchers have thoroughly reviewed supervised ML algorithms, specifically exploring their applications in real-world scenarios. For example, in ecological contexts, Crisci, Ghattas, and Perera (2012) reviewed several well-known supervised ML algorithms, including DT, RF, SVMs, and k-NN. Similarly, Uddin et al. (2019) performed a review of SL algorithms for detecting disease risk. The algorithms include SVM, LR, DT, k-NN, RF, NB, and multi-layer perceptron (MLP). The study recognized the challenge posed by the wide variability in clinical data and research scopes across disease prediction studies. To address this challenge, the study established a common benchmark by selectively choosing studies that implement multiple ML methods on the same dataset and scope for disease prediction.

Belavagi and Muniyal (2016) evaluated four supervised ML algorithms, namely LR, NB, SVM, and RF, for intrusion detection using a popular dataset containing normal and intrusion cases. Patel and Patel (2021) reviewed several SL techniques for predicting agricultural crop yield. These techniques include k-NN, SVM, RF, and NB for predicting a suitable crop for a specific piece of land using data containing seasons and soil parameters. With a focus on supervised ML algorithm applications in banking, Hu et al. (2021) presented an overview of DT, ensemble methods (boosting and bagging), and artificial neural network (ANN), highlighting the importance of these algorithms in the banking sector and demonstrating their utility in a variety of applications, including risk assessment and customer segmentation.

While numerous reviews have extensively covered various domains, the majority exhibit a narrow focus, particularly as seen in references (Nasteski, 2017; Osisanwo et al., 2017; Singh et al., 2016; Uddin et al., 2019). Notably, these reviews have not sufficiently addressed the range of supervised deep-learning algorithms across essential application areas. Our contribution in this context is significant—following

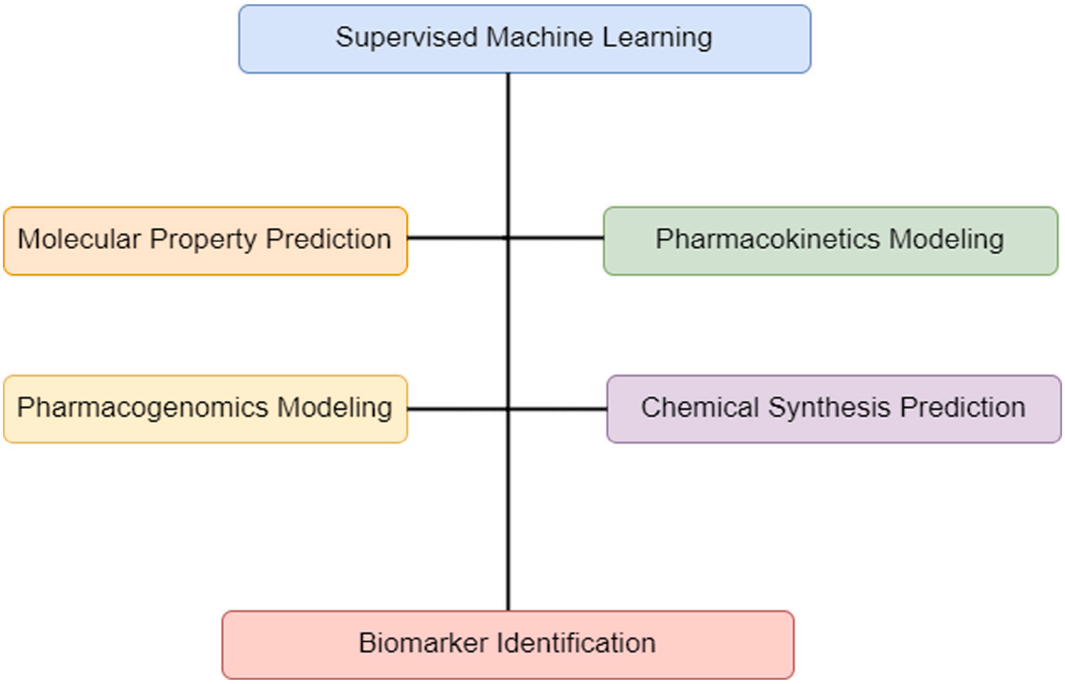
Supervised Machine Learning
Molecular Property Prediction
Pharmacokinetics Modeling
Pharmacogenomics Modeling
Chemical Synthesis Prediction
Biomarker Identification

Fig. 1. Supervised ML-based areas in drug discovery and development.

the emergence of numerous supervised deep learning algorithms, there has been a noticeable lack of comprehensive reviews. Consequently, our review is timely and serves to fill this research gap, providing a more inclusive examination of SL, including supervised deep learning algorithms.

3. Overview of machine learning

ML allows computers to learn by experience and make predictions without being explicitly programmed. It involves using algorithms that can recognize patterns and predict from a given data (Sarker, 2021). There are different ML types based on the training data (as shown in Fig. 2), and they are discussed as follows:

- Supervised learning: This involves using labeled data in training the model (Aruleba et al., 2022). The trained model can then be applied to classify or predict unseen instances. SL can be further split into classification and regression. Classification uses algorithms to predict a categorical label, such as whether a patient is sick or healthy (Agbele, Oriogun, Seluwa, & Aruleba, 2015; Aruleba et al., 2022). The algorithm receives a set of labeled data and learns to classify new data into one of the predefined categories. For example, given a dataset containing patients' records, a classifier can learn to differentiate between the sick and healthy classes and accurately classify them into either class. Meanwhile, regression involves using the algorithm to learn to predict a continuous numerical value, such as a house price. The model learns from a dataset of labeled samples and generates a function that can predict the numerical value of unseen instances (Baştanlar & Özuysal, 2014).
- Unsupervised learning: In this ML type, the algorithm learns and finds patterns in data without prior knowledge of the outcomes. It is used when the data is not labeled (James, Witten, Hastie, Tibshirani, & Taylor, 2023). It can be further divided into two groups: clustering and association. Clustering involves grouping similar data points, while association involves finding relationships between features in the data. Both techniques help identify patterns and gain insights from unstructured data.
- Semi-supervised learning: This ML type incorporates unsupervised and supervised learning methods. It is used when there is insufficient labeled data, and the cost of labeling more data is high. It involves using a small set of labeled data to guide the model in learning the hidden patterns in the data, which is then used in making predictions on unlabeled data. Therefore, the algorithm can learn from both labeled and unlabeled data and gradually enhance its performance. Some widely used approaches include co-training, self-training, and multi-view learning (Kumar, Kaur, & Singh, 2020).
- Reinforcement learning: Reinforcement learning (RL) involves training a model using an approach of rewards and punishments, where the algorithm learns to take actions that maximize the reward over time. The aim of RL is for the agent to learn a strategy that maximizes its long-term reward (Sarker, 2021). It has been applied in several fields, including gaming, robotics, and self-driving cars. Several popular reinforcement learning algorithms have been developed, such as Q-learning, a model-free algorithm that uses a table to store the values of the state–action pairs (see Fig. 2).

4. Supervised learning algorithms

There are different SL algorithms and numerous groupings available in the literature. In this study, the algorithms are categorized into probabilistic, linear, nonlinear, DT, boosting, and deep learning techniques.

4.1. Probabilistic models

4.1.1. Naïve Bayes

NB is a probabilistic algorithm that calculates the likelihood that a given input will belong to a specific class based on prior probabilities and probabilities (Arar & Ayan, 2017). This algorithm assumes that all the features are independent of each other. Gaussian, Bernoulli, and Multinomial NB are three types of Naive Bayes classifiers (Xu, 2018). The Gaussian NB assumes that the input features follow a Gaussian

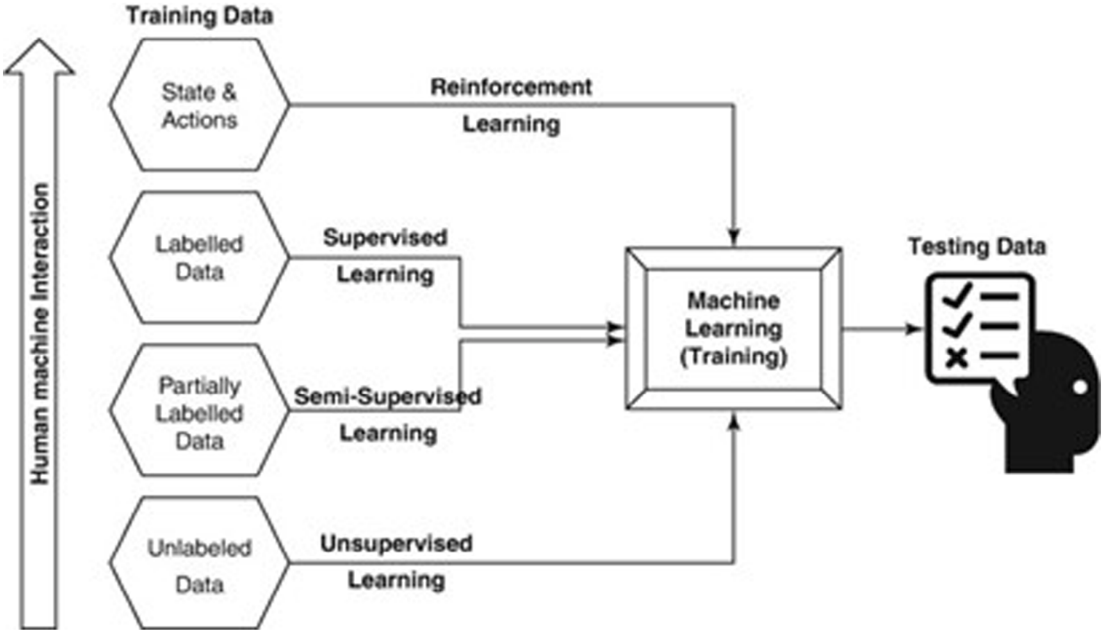

Fig. 2. Types of ML based on training data (El Naqa & Murphy, 2022).

distribution. The Bernoulli NB classifier is used when the input features are binary, such as the existence or absence of a specific attribute, and the Multinomial NB classifier is used when the input features are discrete counts (Singh, Kumar, Gaur, & Tyagi, 2019).

NB is mostly useful when there is a high number of input features and the dataset is sparse. The learning process involves training the classifier with a labeled dataset, where the classifier learns the prior probabilities and conditional probabilities from the training data (Kelly & Johnson, 2021). Meanwhile, Bayes theorem can be represented as:

$$P(Y \mid x_1, x_2, \dots, x_n) = \frac{P(Y)P(x_1, x_2, \dots, x_n \mid Y)}{P(x_1, x_2, \dots, x_n)} \tag{1}$$

where Y represents the class variable, $x_1, x_2, \dots, x_n$ are the independent variables, $P(Y)$, $P(x_i \mid Y)$, $P(x_1, x_2, \dots, x_n \mid Y)$, and $P(x_1, x_2, \dots, x_n)$ represents the prior probability of class Y, the conditional probability of feature x_i, the joint probability of all the features, and the probability of all the features occurring together, respectively. The following formula is employed for predicting the class:

$$\hat{y} = \underset{Y_k}{\operatorname{argmax}} P(Y_k) \prod_{i=1}^{n} P(x_i \mid Y_k) \tag{2}$$

Algorithm 1 summarizes the NB algorithm's learning process, summarizing how the algorithm employs a probabilistic approach to classify a given instance. It starts by calculating the prior probabilities of each class and the conditional probabilities of each feature given a class. For each instance, the algorithm calculates the product of these probabilities for every class, selecting the class with the highest resulting probability as the prediction.

Algorithm 1 Naïve Bayes Algorithm

```
procedure NaiveBayesClassifier($X, Y$)
    Input: $X$, $Y$
    Output: Predicted class labels for a given sample
    Compute prior probabilities $P(Y_k)$ for each class $Y_k$
    for each feature $x_i$ in $X$ do
        Compute conditional probabilities $P(x_i|Y_k)$
    end for
    for each input sample $x = \{x_1, x_2, ..., x_n\}$ in $X$ do
        Initialize $maxProb \leftarrow 0$
        Initialize $predictedClass \leftarrow null$
        for each class $Y_k$ do
            $prob \leftarrow P(Y_k)$
            for each feature $x_i$ in $x$ do
                $prob \leftarrow prob \times P(x_i|Y_k)$
            end for
            if $prob > maxProb$ then
                $maxProb \leftarrow prob$
                $predictedClass \leftarrow Y_k$
            end if
        end for
        Output $predictedClass$ for $x$
    end for
end procedure
```

4.1.2. Bayesian network

The Bayesian network, also called the Bayes network, is a probabilistic graphical classifier that is utilized for probabilistic reasoning and decision-making (Seixas, Zadrozny, Laks, Conci, & Saade, 2014). The Bayesian network represents a collection of random variables and their conditional dependencies using the directed acyclic graph (DAG). Bayesian networks are especially robust when modeling complex systems with numerous variables and intricate interactions between them. The ability of Bayesian networks to provide accurate predictions when given incomplete data is one of its main advantages. Additionally, Bayes networks are simple to update when new data becomes available, which enables them to adjust to evolving conditions.

Bayes networks are valuable classifiers in diverse fields since they can handle incomplete data and still make useful predictions (Kyrimi et al., 2021; Seixas et al., 2014). To perform probabilistic inference in a Bayesian network, we can use Bayes' rule, which in this case can be written as follows:

$$P(X_i \mid \text{evidence}) = \frac{P(\text{evidence} \mid X_i)P(X_i)}{\sum_{X_i} P(\text{evidence} \mid X_i)P(X_i \mid \text{parents}(X_i))}. \tag{3}$$

where $P(\text{evidence} \mid X_i)$ is the likelihood of the evidence given X_i, $P(X_i)$ is the prior probability of X_i, and the denominator is the normalization constant.

4.2. Linear classifiers

Linear classifiers are a type of SL algorithm that creates a linear boundary between classes to classify the input data. The most common type of linear classifier is the perceptron algorithm. Other examples include logistic regression and SVM. This category of SL algorithms is known for their simplicity and efficiency and is often used in applications where interpretability and speed are important factors.

4.2.1. Logistic regression

LR is a statistical method used for predicting binary outcomes (Harris, 2021). It is a type of regression analysis where the outcome variable is categorical and takes only two values: 0 or 1. Given a set of input variables, it is used to model the probability that a particular event occurs (Austin & van Buuren, 2023; van Smeden, Moons, de Groot, Collins, Altman, Eijkemans, & Reitsma, 2019). Algorithm 2 describes and summarizes the learning process of LR, including parameter initialization and iterative optimization of the cost function, which is based on the logistic (i.e. sigmoid) function.

Algorithm 2 Logistic Regression

```
procedure LogisticRegressionTraining($X, Y, \alpha$, iterations)
    Input:
        $X$ - feature set (input variables)
        $Y$ - target class labels (output variable)
        $\alpha$ - learning rate
        iterations - number of training iterations
    Output: Model parameters $\Theta$
    Initialize model parameters $\Theta$ with zeros
    for $i \leftarrow 1$ to iterations do
        Compute the hypothesis $h_\Theta(x) = \frac{1}{1+e^{-\Theta^T X}}$
        Compute the cost $J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(h_\Theta(x^{(i)})) + (1 - y^{(i)})\log(1 - h_\Theta(x^{(i)}))]$
        Compute the gradient $\nabla J(\Theta) = \frac{1}{m} X^T(h_\Theta(X) - Y)$
        Update parameters $\Theta \leftarrow \Theta - \alpha\nabla J(\Theta)$
    end for
    return $\Theta$
end procedure
```

The algorithm employs the gradient descent optimization technique to minimize the cost function of the model. As shown in Algorithm 2, the model parameters are adjusted iteratively based on the difference between the predicted outcome and the actual class labels. Furthermore, it involves determining the cost function's gradient, which quantifies the prediction error over the input data, and updating the parameters in the direction that reduces this error. The learning rate α controls the size of the parameter updates, and the procedure is repeated for a predefined number of iterations.

4.2.2. Linear discriminant analysis

Linear Discriminant Analysis (LDA) is commonly used for dimensionality reduction and classification tasks, especially when the classes are well-separated, and the assumptions of LDA are met. The goal of LDA is to project the data onto a lower-dimensional subspace while preserving the class-discriminatory information as much as possible (Chen et al., 2019; Seng & Ang, 2017). If X is the $N \times D$ matrix representing N samples with D features, and y be the corresponding class labels. The class means μ_k and within-class scatter matrices S_w are calculated as follows:

$$mu_k = \frac{1}{N_k}\sum_{i=1}^{N} x_i, \quad S_w = \sum_{k=1}^{K}\sum_{i=1}^{N_k}(x_i - \mu_k)(x_i - \mu_k)^T \tag{4}$$

where N_k is the number of samples in class k. The between-class scatter matrix S_b is given by:

$$S_b = \sum_{k=1}^{K} N_k(\mu_k - \mu)(\mu_k - \mu)^T \tag{5}$$

where μ is the overall mean of all classes. Furthermore, the LDA aims to find the projection matrix W that maximizes the ratio of between-class scatter to within-class scatter, given by:

$$max_W \frac{\text{tr}(W^T S_b W)}{\text{tr}(W^T S_w W)} \tag{6}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix. The optimal projection matrix W is the one corresponding to the largest eigenvalues of $S_w^{-1}S_b$. LDA assumes that the classes have a Gaussian distribution with the same covariance matrix and that the classes are linearly separable (Thomaz, Kitani, & Gillies, 2006). When these assumptions hold, LDA provides a simple and effective classification method.

4.3. Nonlinear classifiers

4.3.1. Support vector machines

SVMs can be used for linear and non-linear classification tasks, making them highly adaptable to various real-world applications (Gholami & Fakhari, 2017). The algorithm uses the concept of a hyperplane, which is a decision boundary that separates data points into different classes. The goal of SVMs is to find the hyperplane that maximizes the margin. This margin maximization approach allows SVMs to achieve good generalization performance and handle data that may not be linearly separable (Suthaharan & Suthaharan, 2016). The optimization task can be formulated as:

$$\min_{w,b}\frac{1}{2}\|w\|^2 \quad \text{s.t. } y_i(w^N x_i + b) \geq 1 \quad \forall i = 1, \dots, n \tag{7}$$

where w represents the weight vector, b is the bias term, and $\|w\|$ is the Euclidean norm of w. Meanwhile, w and b can be computed using:

$$w = \sum_{i=1}^{n}\alpha_i y_i x_i \quad \text{and} \quad b = y_k - \sum_{i=1}^{n}\alpha_i y_i x_i^T x_k \tag{8}$$

where k is any support vector with $\alpha_k > 0$.

4.3.2. K-Nearest Neighbors

k-NN is a simple and intuitive non-parametric classification algorithm. The main idea behind k-NN is to classify a new data point based on the majority class of its nearest neighbors in the feature space (Anava & Levy, 2016). The distance metric plays a crucial role in determining the neighbors. While the Euclidean distance is commonly used, k-NN can also utilize non-Euclidean distance metrics to capture the underlying structure of the data better.

Assuming x_i represents a data point in the feature space, and x_j denotes its nearest neighbor. The non-Euclidean distance $d(x_i, x_j)$ between x_i and x_j can be calculated using various metrics such as Manhattan distance, Minkowski distance, or Mahalanobis distance, depending on the characteristics of the data. For example, the Manhattan distance between two data points x_i and x_j in n-dimensional space is defined as:

$$d(x_i, x_j) = \sum_{k=1}^{n}|x_{ik} - x_{jk}| \tag{9}$$

where x_{ik} and x_{jk} are the k-th features of x_i and x_j, respectively. This distance metric is particularly useful when dealing with high-dimensional data.

4.4. *Decision trees*

DT is a well-known ML algorithm that creates a tree-like structure of decisions and their possible outcomes. At every node of the tree, a decision is reached based on specific criteria determined by analyzing the data. The tree comprises nodes, branches, and leaves that represent decisions, possible outcomes of those decisions, and the final outcome of the decision process (Gavankar & Sawarkar, 2017). DT are known for their interpretability; the tree-like structure allows researchers to easily understand and visualize the decision-making process, which is critical when validating models for regulatory approval (Mienye & Jere, 2024). DT are also versatile and can handle both numerical and categorical data, making them suitable for tasks such as predicting drug-target interactions, classifying compounds based on their biological activity, and identifying key molecular features that contribute to drug efficacy. Meanwhile, there are different types of DTs, including:

4.4.1. *Classification and regression trees*

The classification and regression tree (CART) algorithm is used for classification and regression. The algorithm selects the input variable that provides the best split. The best split is defined as the one that maximizes the difference between the parent node's impurity and the weighted impurity of the child nodes. The impurity of a node is a measure of how mixed the class or target variable values are within that node. The impurity measure used for classification tasks is typically Gini impurity (Breiman, 2017). Algorithm 3 describes the CART Algorithm. At each node, CART selects the split that produces the purest child nodes, continuing this process until it meets stopping criteria, such as a maximum tree depth or a minimum node size. The algorithm then prunes the tree to mitigate overfitting, enhancing its generalization capabilities.

Algorithm 3 CART Algorithm

```
Input: $X$, $Y$, maximum depth of the tree (max_depth), minimum size of a node (min_size)
Output: CART-based DT model
Procedure
    Initialize tree
    Split the root node based on the best-split point
    Recursively split child nodes
    Stop if maximum depth or minimum node size is reached
    Prune the tree
return DT model
End Procedure
```

4.4.2. *Iterative Dichotomiser 3*

The Iterative Dichotomiser 3 (ID3) algorithm is described in Algorithm 4. It selects features to split the data based on the Information Gain (IG) criterion, with the goal of maximizing the reduction in entropy (Mienye et al., 2019).

Algorithm 4 begins with the entire dataset and evaluates the ability of each attribute to classify the data effectively, selecting the attribute that results in the highest IG for each split. The process continues recursively until all data is perfectly classified.

4.4.3. *C4.5*

The C4.5 DT (Algorithm 5) is an extension and improvement of the ID3 algorithm. Some of the improvements include its ability to hand both continuous and categorical data in classification tasks. C4.5 employs the information gain ratio to select the most informative feature at each node and also incorporates mechanisms to handle missing values, tree pruning to avoid overfitting (Mienye et al., 2019). It handles continuous attributes by dynamically defining threshold values for splits. The C4.5 iteratively partitions the given data until each subset is pure or predefined stopping criteria are met, leading to DT models that can effectively classify new instances.

Algorithm 4 ID3 Algorithm

```
Input: $X$, $Y$
Output: ID3-based DT model
Procedure
if All samples are in the same class then
    return leaf node with class label
end if
if No features left to split then
    return leaf node with the most common class label
end if
Select feature with highest IG as node
Split dataset based on feature values
Recursively apply ID3 to each subset
return DT model
End Procedure
```

Algorithm 5 C4.5 Algorithm

```
 1: Input: $X$, $Y$, stopping criteria (thresholds)
 2: Output: DT model
 3: Procedure
 4: if All examples are in the same class or other stopping criteria met then
 5:     return leaf node with class label
 6: end if
 7: Select feature with highest information gain ratio
 8: Split dataset based on feature values or threshold for continuous data
 9: Handle missing values
10: Recursively apply the splitting and selection steps (steps 5-9) to each subset
11: Apply pruning to reduce tree size and complexity
12: return  C4.5 model
13: End Procedure
```

4.4.4. *Random forest*

RF is an ensemble algorithm that combines multiple DT models to create a more powerful model. This algorithm is known for its ability to model complex data and provide reliable predictions. The algorithm works by creating many DT and aggregating their predictions (Obaido et al., 2024; Zou & Schonlau, 2018). The different DT are trained using a random subset of the input data and a random subset of the given features. This random approach ensures overfitting is reduced and improves the final model's generalization performance (Mienye & Sun, 2022). The algorithm takes the average prediction of the base models to predict new unseen instances. The RF algorithm can deal with categorical and continuous variables, and unlike single DT models, it is less prone to overfitting. Given an instance x with F classes, the final ensemble prediction from N trees can be calculated as follows:

$$H(N(x)) = \arg\max_{j} \sum_{k=1}^{K} \mathbf{1}(h_k(x) = j), \text{ for } j = 1, \dots, C \tag{10}$$

4.5. *Boosting*

Boosting is an ensemble technique that iteratively improves the performance of individual models by giving more weight to misclassified instances in subsequent iterations, ensuring the models learn from their mistakes and make better predictions.

4.5.1. *XGBoost*

The XGBoost is an implementation of gradient boosting that has been optimized for speed and performance (He, Hao, & Wang, 2021). The XGBoost algorithm iteratively builds multiple base models, where new models attempt to correct the errors of the preceding models. The final prediction is then computed through the summation of all the base model's predictions (Cheong et al., 2021; Dhaliwal, Nahid, & Abbas, 2018). The XGBoost's objective function is represented as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{11}$$

where θ denotes the model parameters, n is the number of instances, y_i and $\hat{y}_i$ represents the actual label and predicted label of the ith instance, $l(y_i, \hat{y}_i)$ is the loss function, K is the number of trees, and $\Omega(f_k)$ is the regularization term (Li & Chen, 2020). At every iteration, XGBoost fits a new model to correct the errors of the previous model, which is achieved by minimizing the objective function using the gradient descent algorithm.

4.5.2. AdaBoost

The adaptive boosting (AdaBoost), similar to other boosting algorithms, combines multiple weak classifiers to form a strong classifier (Cui, Chen, Wang, Li, & Ling, 2021; Mienye, Obaido, Aruleba, & Dada, 2021). It focuses on the misclassified data points and iteratively trains the model to improve its performance. The algorithm starts by assigning equal weights to every training example. Let D_1 be the weight vector for the first round of training, where $D_{1,i} = \frac{1}{n}$, at every iteration, a weak classifier $h_j(x)$ is trained using the weighted training data. The weak classifier is trained to minimize the weighted error rate E_j:

$$E_j = \sum_{i=1}^{n} D_{j,i} I(y_i \neq h_j(x_i)) \tag{12}$$

where y_i and x_i are the true label and feature vector of the instance i, and I is the indicator function. The weight α_j of the weak classifier is then computed as

$$\alpha_j = \frac{1}{2} \ln \frac{1 - E_j}{E_j} \tag{13}$$

The weights of the training samples are updated based on the performance of the weak classifier (Zheng, Xiao, Sun, & Qin, 2022). The weight of data point i in the $(j+1)^{th}$ round, $D_{j+1,i}$, is computed as

$$D_{j+1,i} = \frac{D_{j,i} \exp(-\alpha_j y_i h_j(x_i))}{Z_j} \tag{14}$$

where Z_j is the normalization factor and $Z_j = \sum_{i=1}^{n} D_{j,i} \exp(-\alpha_j y_i h_j(x_i))$. The purpose of the weight update is to give more weight to the misclassified instances (Sevinç, 2022). Assuming we have N total number of weak classifiers, the final classification model is the weighted combination of the base models:

$$H(x) = sign\left(\sum_{j=1}^{N} \alpha_j h_j(x)\right) \tag{15}$$

AdaBoost has significant advantages in drug discovery by combining multiple weak classifiers to form a strong classifier, which improves the predictive performance. Furthermore, AdaBoost excels in scenarios where the primary challenge is class imbalance, a common issue in drug discovery datasets where the number of active compounds is often much smaller than inactive ones. By focusing on misclassified instances, AdaBoost enhances the model's ability to detect minority class instances, such as rare but potentially highly effective compounds. This characteristic makes AdaBoost particularly valuable in early-stage drug discovery, where identifying novel active compounds is crucial.

4.5.3. CatBoost

Categorical boosting (CatBoost) is a popular ensemble learning method developed by Prokhorenkova, Gusev, Vorobev, Dorogush, and Gulin (2018) to combine multiple weak learners to obtain a strong ensemble classifier. CatBoost is known for handling categorical features without needing one-hot encoding. It is also able to handle missing data and has built-in text data support, making it suitable for natural language processing. It employs a symmetric tree structure for its DTs, which helps to reduce overfitting. Furthermore, CatBoost can handle imbalanced data using a specialized objective function that considers the class distribution in the data. The objective function is defined as follows:

$$F = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{i=1}^{N} \Omega(f_i) \tag{16}$$

$L(y_i, \hat{y}_i)$ is the loss function, f_i is the ith tree, and $\Omega(f_i)$ denotes the regularization term that penalizes complex trees. This algorithm uses the logarithmic loss function, and its regularization term is the L2 regularization, which is represented as follows:

$$L(y_i, \hat{y}_i) = -(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)) \tag{17}$$

$$\Omega(f_i) = \frac{1}{2} \lambda \|w\|^2 \tag{18}$$

where w represents the vector of weights and λ is the regularization parameter. The CatBoost algorithm uses the gradient boosting technique to optimize its cost function. The CatBoost algorithm also uses a method known as ordered boosting to deal with categorical variables. Ordered boosting is a variant of gradient boosting that orders categorical features based on their contribution to the objective function. The ordered features are then used to split the data, resulting in better performance and faster convergence.

4.5.4. LightGBM

The Light Gradient Boosting Machine (LightGBM), developed by Microsoft in 2017, is popular for its efficiency and speed, especially when working with large datasets (Ke et al., 2017). LightGBM is suitable for different machine-learning tasks, including regression and classification. Like other boosting algorithms, the LightGBM training process involves iteratively adding DTs to an ensemble, where the trees aim to correct the classification errors from the preceding trees. The process starts with a single DT, which is trained on the entire dataset. The subsequent trees are then trained on the misclassification of these preceding trees, thereby enhancing the overall model's classification performance. LightGBM builds a DT by minimizing the objective function E:

$$E(y, F(x)) = \sum_{i=1}^{n} l(y_i, F_i(x_i)) + \Omega(F) \tag{19}$$

where $F_i(x_i)$ is the predicted value of x_i, $l(y_i, F_i(x_i))$ and $\Omega(F)$ denotes the loss function and the regularization term, respectively. LightGBM uses a histogram-based algorithm to split the data into bins, reducing memory usage and ensuring a fast training process. The histogram-based algorithm works as follows:

1. For each feature, sort the feature values in ascending order.
2. Divide the sorted feature values into discrete bins.
3. Calculate the histogram of the labels for each bin.
4. Find the best-split point based on the histogram information gain, which is calculated as:

$$IG = \frac{1}{2}\left(\frac{GL^2}{HL + \lambda} + \frac{GR^2}{HR + \lambda} - \frac{G^2}{H + \lambda}\right) \tag{20}$$

where GL and HL are the sum of labels and weights in the left child node, GR and HR are the sum of labels and weights in the right child node, G and H are the sums of labels and weights in the current node, and λ is the regularization parameter.

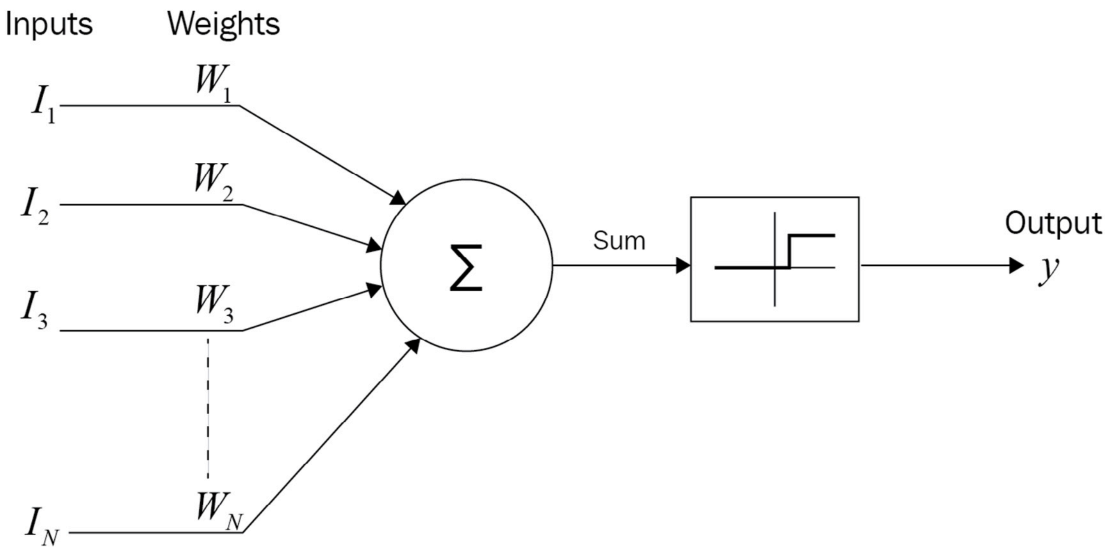

Fig. 3. McCulloch–Pitts neuron model.

4.6. *Deep learning*

Deep learning-based approaches use multiple layers of interconnected nodes to analyze complex non-linear relationships between the predictor and response variables. Supervised deep learning has shown remarkable success and is widely used in numerous real-world applications (Deng & Li, 2013; Hirschberg & Manning, 2015; Liu et al., 2017; Rothman, 2018; Scarselli, Gori, Tsoi, Hagenbuchner, & Monfardini, 2008; Shinde & Shah, 2018; Wu, Sun, Zhang, Xie, & Cui, 2022). This section describes some SL approaches. However, before going into the deep learning architectures, it is necessary to introduce the main building block and core foundation of deep learning, i.e., the ANN. The ANN is influenced by how biological nervous systems, like the brain, process data (Dongare, Kharde, Kachare, et al., 2012; Kukreja, Bharath, Siddesh, & Kuldeep, 2016; Mienye & Sun, 2023). It consists of numerous intricately linked processing units or neurons that collaboratively address a particular challenge. It can be visualized as a weighted directed graph comprising neurons and directed edges, each with its associated weight (Jain, Mao, & Mohiuddin, 1996).

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \tag{21}$$

where y represents the output matrix and f denotes the activation function applied to each element. Furthermore, w, x, and b represent the weight matrix, input matrix, and bias vector. One of the earliest and foundational models of a neuron in ANN literature is the McCulloch–Pitts neuron (Dybowski & Gant, 2001; Rothman, 2018). It takes multiple binary inputs, multiplies each with a weight, and then produces a binary output based on a threshold activation function, as described in Fig. 3. ANNs can be categorized into *feed-forward networks*, which are characterized by graphs without loops, and *feedback* or *recurrent networks*, distinguished by the loops resulting from their feedback connections. A typical ANN is mathematically represented as follows:

4.6.1. *Convolutional neural networks*

A convolutional neural network (CNN) is a specialized class of neural networks designed to process data using three fundamental building blocks or layers (Bentes, Velotto, & Tings, 2017; Kattenborn, Leitloff, Schiefer, & Hinz, 2021). These networks are particularly renowned for their proficiency in image and video recognition tasks. In a typical CNN, as shown in Fig. 4, the first two layers, the convolution and pooling layers, excel at extracting relevant features from the given data. Given the training data, the CNN model undergoes forward propagation, where the specific kernels and weights contribute to a loss function calculation. These learnable parameters are then updated through backpropagation using the gradient descent optimization technique based on the computed loss value. After the convolution and pooling layers, the output is usually flattened and fed into one or more fully connected (FC) layers. The CNN output y can be computed as:

$$y = \sigma(W \cdot X + b)$$

where n represents the number of neurons in the FC layer, σ is the activation function, X indicates the input to the FC layer, W is the weight matrix, and b is the bias vector.

CNNs are particularly effective in processing spatial data, making them ideal for analyzing high-content screening images and predicting molecular properties from chemical structures. Their ability to automatically extract hierarchical features from input data makes CNNs a robust algorithm for identifying biologically relevant patterns without extensive manual feature engineering.

4.6.2. *Recurrent neural networks*

A recurrent neural network (RNN) is a type of ANN that is built for processing sequential data. It is characterized by its ability to retain information from previous inputs and use it to make predictions. The RNN architecture consists of a series of interconnected nodes that are organized in a directed cycle, allowing information to flow in a loop (Dernoncourt, Lee, Uzuner, & Szolovits, 2017; Schuster & Paliwal, 1997). Unlike feedforward neural networks, which process each input independently, recurrent neural networks have a memory component that allows them to retain information from previous inputs. Also, in the RNN, shown in Fig. 5, each neuron receives input not only from the current time step but also from the output of the previous time step, creating a recurrent connection. This loop-like structure enables RNNs to retain the memory of previous inputs and learn dependencies within the sequential data.

The Long short-term memory (LSTM) and gated recurrent unit (GRU) are advanced variants of the RNN, which were developed to solve the vanishing gradient problem associated with the simple RNN.

Furthermore, Simple RNN and advanced variants like LSTM and GRU perform well at handling sequential data, which is common in pharmacokinetics and pharmacodynamics modeling. These models can capture temporal dependencies and predict future outcomes based on historical data, making them valuable for modeling drug response over time and optimizing dosage regimens. The memory retention capabilities of RNNs are crucial for understanding and predicting the dynamic interactions of drugs within biological systems.

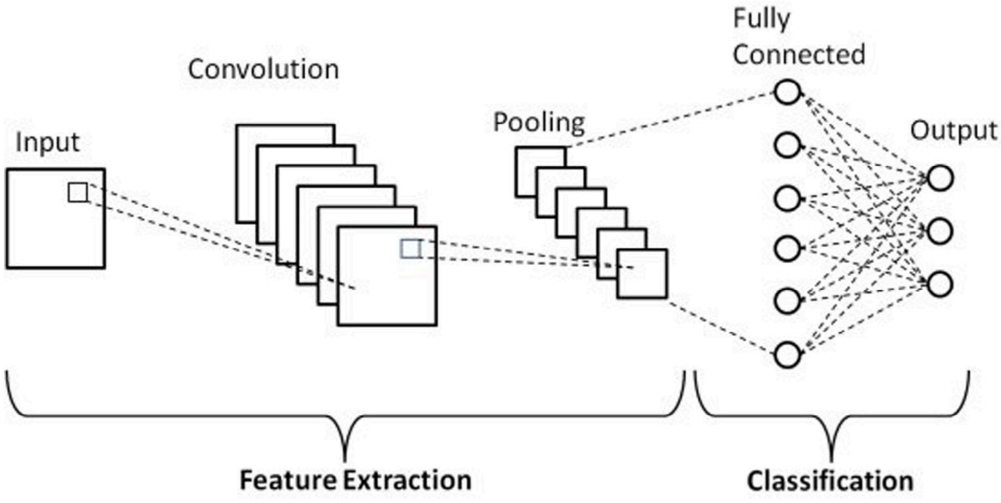

Fig. 4. Basic CNN architecture.

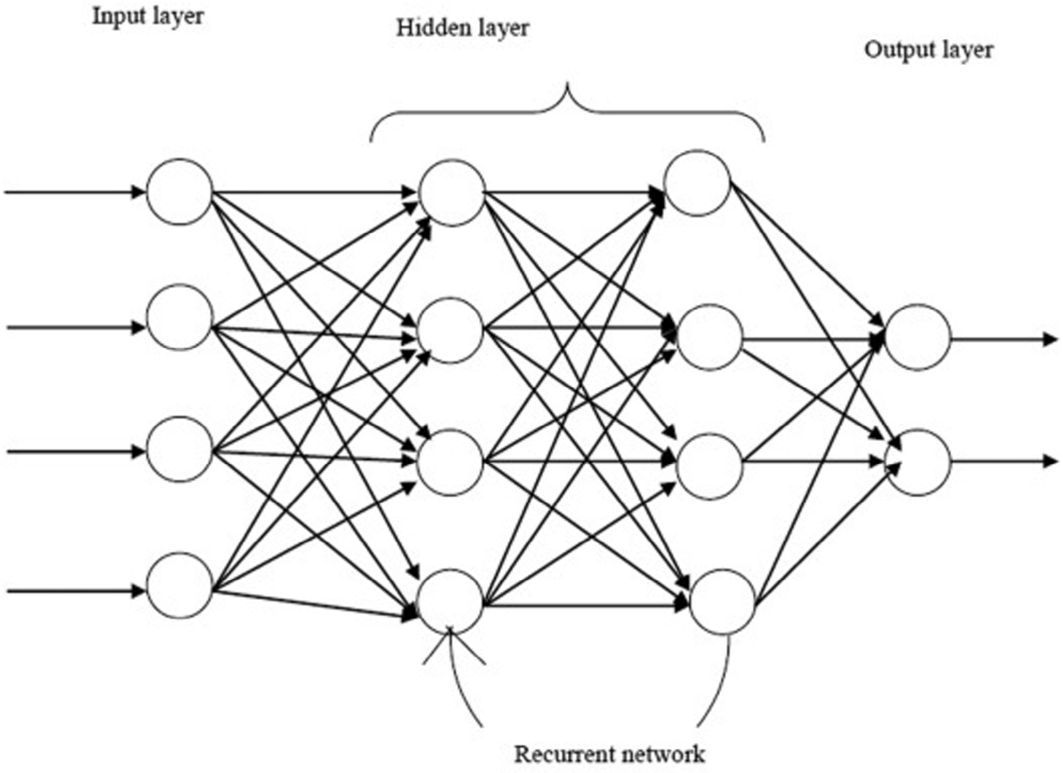

Fig. 5. A RNN architecture.

4.6.3. *Graph neural networks*

A graph neural network (GNN) is designed to perform operations on data described by graphs (Scarselli et al., 2008; Wu et al., 2022). GNNS are specifically tailored to process data with complex dependencies and relationships represented as graphs, where nodes are connected by edges. GNNs aim to learn representations for each node in the graph, incorporating information from neighboring nodes and their relationships. The pipeline of a GNN model, as described in Fig. 6, involves iteratively aggregating information from neighboring nodes, updating node embeddings, and repeating the process for multiple layers to capture complex dependencies and information diffusion across the graph.

$$m_v = \sum_{u \in N(v)} f(h_u, e_{vu}) \tag{22}$$

where m_v is the aggregated message for node v, obtained by combining information from its neighboring nodes $N(v)$ using the aggregation function f. The aggregation function can take into account both the hidden representations of the neighboring nodes h_u and the edge features e_{vu}.

GNNs are uniquely suited to modeling data that can be represented as graphs, such as molecular structures and protein–protein interaction networks. By aggregating information from connected nodes, GNNs can capture the intricate dependencies and relationships within the data. This capability is particularly beneficial for predicting molecular activity, identifying potential drug targets, and exploring the interactions within biological pathways. The ability of GNNs to handle complex relational data makes them indispensable in the computational analysis of molecular and genetic data in drug discovery.

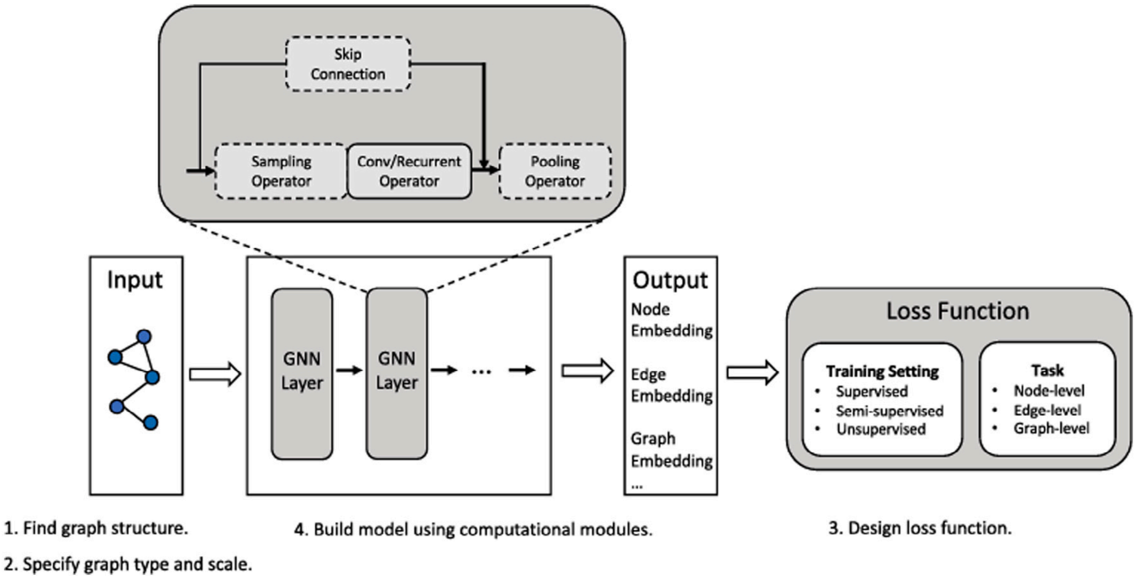

Fig. 6. The pipeline of a GNN model (Zhou et al., 2020).

Lastly, deep learning techniques, such as CNNs, RNNs, and GNNs, offer significant advantages in drug discovery due to their ability to learn complex patterns from vast and diverse datasets. By integrating these DL techniques, researchers can leverage the strengths of each algorithm to address specific challenges in drug discovery. For example, combining CNNs with GNNs can enhance the prediction of molecular properties by incorporating both spatial and relational information, leading to more accurate and robust models. Similarly, integrating RNNs with GNNs can improve the modeling of temporal dynamics in biological networks, facilitating the discovery of time-dependent drug interactions and effects. The synergistic use of these deep learning approaches provides a comprehensive toolkit for advancing drug discovery and development.

5. Common performance metrics

In supervised ML, performance metrics play a vital role in assessing the effectiveness of models and guiding improvements. These metrics can be broadly categorized into those used for classification tasks and regression tasks (Botchkarev, 2018; Cuadros-Rodríguez, Pérez-Castaño, & Ruiz-Samblás, 2016; Koyejo, Natarajan, Ravikumar, & Dhillon, 2014).

5.1. Classification metrics

Classification metrics provide a means to evaluate the performance of a classification algorithm (Mienye & Jere, 2024; Zhou, Gandomi, Chen, & Holzinger, 2021). These metrics provide an understanding of how well a model is performing in distinguishing between classes. These metrics include accuracy, precision, sensitivity, specificity, and F1-Measure.

$$Accuracy\ (Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

$$Sensitivity\ (Sen) = \frac{TP}{TP + FN} \tag{25}$$

$$Specificity\ (Spec) = \frac{TN}{TN + FP} \tag{26}$$

$$F1\ measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{27}$$

$$True\ Positive\ Rate = \frac{TP}{TP + FN} \tag{28}$$

$$False\ Positive\ Rate = \frac{FP}{FP + TN} \tag{29}$$

Where:

- True Positives (TP) represent the number of correctly predicted positive instances.
- False Negatives (FN) represent the number of positive instances incorrectly classified as negative.
- True Negatives (TN) represent the number of correctly predicted negative instances.
- False Positives (FP) represent the number of negative instances incorrectly classified as positive.
- True Positive Rate (TPR) measures the proportion of actual positives that are correctly identified by the model.
- False Positive Rate (FPR) measures the proportion of actual negatives that are incorrectly identified as positives by the model. It highlights the rate at which false alarms occur.

AUC (Area Under the Curve) is a performance metric used to evaluate the ability of a binary classification model to distinguish between positive and negative classes. Specifically, it is associated with the ROC (Receiver Operating Characteristic) curve, which plots the TPR against the FPR at various threshold settings. An AUC of 0.5 indicates no discrimination capability, equivalent to random guessing. An AUC of 1 signifies perfect discrimination capability. An AUC below 0.5 suggests the model is worse than random guessing, consistently misclassifying classes. Values between 0.9 and 1 indicate excellent performance, 0.8 to 0.9 good performance, 0.7 to 0.8 fair performance, 0.6 to 0.7 poor performance, and 0.5 to 0.6 indicate the model fails to perform adequately.

5.2. Regression metrics

Regression metrics provide a way to evaluate the performance of a regression model, which predicts a continuous outcome (Naidu, Zuva, & Sibanda, 2023; Yildiz, Bilbao, & Sproul, 2017). In this section, we provide some metrics used for regression analysis.

Mean Absolute Error (MAE): MAE is the average of the absolute differences between predicted and actual values. It provides a straightforward measure of prediction accuracy.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{30}$$

Mean Squared Error (MSE): MSE is the average of the squares of the differences between predicted and actual values. Squaring the errors places more weight on larger errors, making MSE sensitive to outliers.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{31}$$

Root Mean Squared Error (RMSE): RMSE is the square root of MSE, translating the error into the same units as the original values. It also emphasizes larger errors more than MAE.

$$RMSE = \sqrt{MSE} \tag{32}$$

R-squared (R^2): R^2 measures the proportion of variance in the dependent variable that is predictable from the independent variables. It indicates the goodness of fit of the model. However, it does not account for overfitting.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}i)^2}{\sum i = 1^N (y_i - \bar{y})^2} \tag{33}$$

Adjusted R-squared: This metric adjusts R^2 for the number of predictors in the model, providing a more accurate measure in the presence of multiple predictors.

$$Adjusted\ \ R^2 = 1 - \left(\frac{1 - R^2}{N - k - 1} \right) \times (N - 1) \tag{34}$$

Correlation (Corr R) is a statistical measure that describes the degree to which two variables move in relation to each other. It is commonly used in regression analysis to measure the strength and direction of the linear relationship between two variables.

$$Corr\ \ R = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{35}$$

Where x_i are individual values of the independent variable x. y_i are the Individual values of the dependent variable y. $\bar{x}$ is the mean of x values, and $\bar{y}$ is the mean of y values. This equation calculates the Pearson correlation coefficient r, which measures the linear relationship between two variables x and y. It quantifies the strength and direction of their linear association, ranging from −1 (perfect negative correlation) to +1 (perfect positive correlation).

6. Applications of supervised learning to drug discovery and development

Several studies have applied SL methods to advance the field of drug discovery and development. These applications span a broad spectrum of processes, ranging from predicting molecular properties to exploring genomics.

Meanwhile, in evaluating the performance of the models, this study focuses on two widely used metrics: Acc and AUC. Accuracy is one of the most straightforward metrics for evaluating the performance of a classification model. It is defined as the ratio of correctly predicted instances to the total instances in the dataset. The AUC is derived from the receiver operating characteristic (ROC) curve, which plots the TPR against the FPR at various threshold settings. The TPR, also known as sensitivity or recall, is the proportion of actual positives correctly identified by the model, while the FPR is the proportion of actual negatives incorrectly identified as positives. The AUC represents the probability that a randomly chosen positive instance is ranked higher by the classifier than a randomly chosen negative instance. An AUC value ranges from 0 to 1, with 0.5 equivalent to random guessing, 1 indicating that the model perfectly separates positive and negative instances, and 0 indicating a perfectly incorrect model, i.e., all predictions are wrong.

6.1. Molecular property and activity prediction

Supervised ML models have been used to predict various molecular properties critical for assessing chemical compound viability as therapeutic agents. Fig. 7 presents a supervised ML technique for molecular property prediction. Tayyebi et al. (2023) applied RF for the prediction of chemical solubility in water using molecular descriptors and Morgan fingerprint methods. The study demonstrated the efficacy of ML techniques in predicting the solubility of various chemical species. Marchetti et al. (2021) used LR, SVM, and RF algorithms to develop a classification model for predicting the functional effects (activation or inhibition) of allosteric ligands on the molecular chaperone Hsp90. Using integrated molecular docking-based screening, protein conformational dynamics information, and ML techniques to train the model on data from ensemble docking results of 133 known Hsp90 ligands.

Zhang et al. (2019) explored the effectiveness of eight ML methods, including DT, k-NN, and several other supervised ML models, in molecular drug design and discovery, particularly focusing on accelerating Acetyl-CoA Carboxylases inhibitors discovery. Feinberg et al. (2018) designed PotentialNet, a family of GNN tailored for predicting molecular properties relevant to drug discovery, particularly focusing on protein–ligand binding affinity. Through leveraging feature learning, the GNN aim to surpass traditional physics-based and knowledge-based ML models in performance.

For several molecular property predictions, Wang et al. (2022) created AdvProp, a tool based on a combination of graph-based and sequence-based supervised ML methods for the molecular property prediction, aimed to accelerate drug discovery efforts, especially for global health pandemic situation, such as COVID-19. For several molecular activity predictors and properties, including but not limited to bioactivity, toxicity, or interactions with specific biological targets, Lane et al. (2020) applied several supervised ML algorithms using SVM classification, NB, AdaBoost, and others for drug discovery applications. The study used the ChEMBL database of 5000 datasets and 570,000 unique combinations generated using the ECFP6 fingerprint. Ashraf, Akter, Mumu, Islam, and Uddin (2023) developed several computational models using LR and several supervised ML models for the prediction of drug potency compounds against the SARS-CoV-2 3CLpro protein, a key therapeutic target for COVID-19 treatment. Furthermore, the study utilized SHapley Additive exPlanations (SHAP) analysis to identify important descriptors for predicting bioactivity. Wallach et al. (2015) developed AtomNet, a CNN model designed specifically for predicting the bioactivity of small molecules in structure-based drug discovery. By hierarchically composing local features, AtomNet demonstrates superior predictive performance, outperforming previous docking approaches on diverse benchmarks. Aly and Alotaibi (2023) explored the application of deep learning, particularly LSTM networks, in predicting the molecular properties of modified Gedunin, a compound of interest in drug research. According to the study, the proposed model achieved a high accuracy in predicting molecular properties, indicating promising potential for rational drug design and exploration. Ahmad, Tayara, Shim, and Chong (2024) created SolPredictor based on the residual GNN convolution (RGNN) for predicting molecular solubility. Overall, SolPredictor demonstrated significant potential in enhancing solubility prediction accuracy, thereby contributing to more efficient and streamlined drug development processes.

Table 1 presents various supervised ML models applied to predict molecular properties and activities. These models include traditional methods such as RF, LR, and SVM, as well as advanced techniques such as Gradient Boosting, XGBoost, and GNN. The models are employed for various predictive tasks, including chemical solubility, molecular ligand classification, compound property identification, protein–ligand binding affinity, and bioactivity prediction. The studies rely on various open databases, such as Vermeire, Boobier, Delaney, the Protein Data Bank, ChEMBL, and PubChem, providing rich datasets for model training and validation. In addition, SHAP was used to provide interpretability, offering insights into feature importance and model decisions.

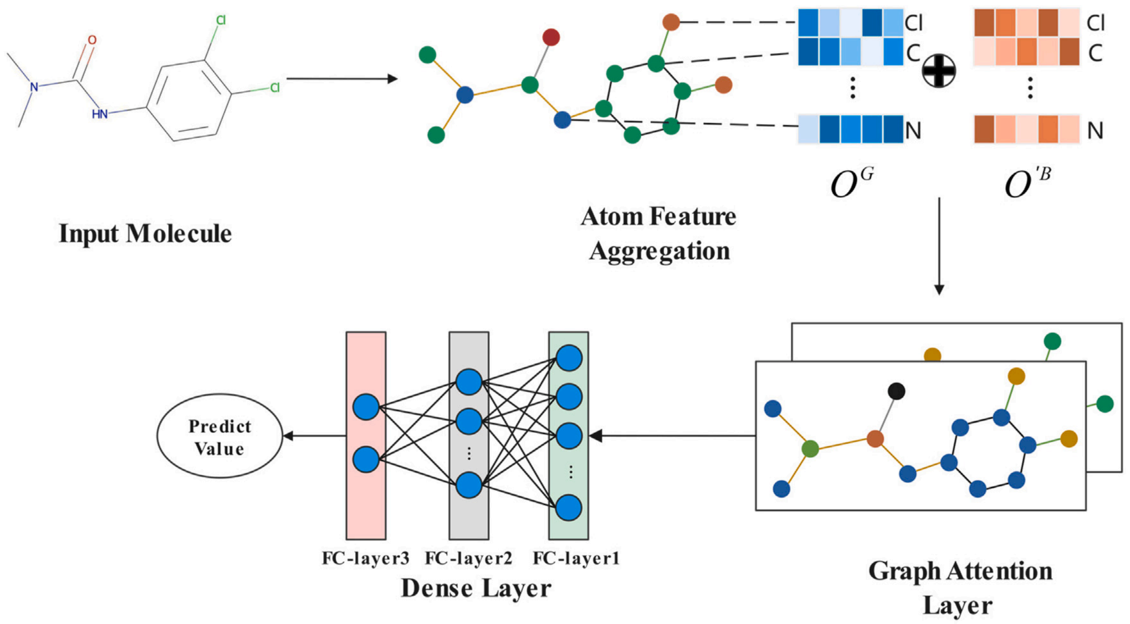

Fig. 7. Structure of a GNN applied for molecular property prediction (Xu, Pan, Xia, & Li, 2023).

Table 1
Summary of supervised ML algorithms used for Molecular Property and Activity Prediction.

Reference	Model	Scope	Performance	Database
Tayyebi et al. (2023)	RF and Shapley Additive exPlanations (SHAP)	Predict chemical solubility	Acc = 88%	Open databases: Vermeire, Boobier and Delaney.
Marchetti et al. (2021)	LR, RF, SVM	Classify molecular ligands	Highest Acc = 89%	Open database: Protein Data Bank
Zhang et al. (2019)	DT, k-NN, SVM, RF, AdaBoost, GB, XGBoost, XT	Identify active or inactive compound property	Highest Acc = 89.5%	Open database: Crystal Protein Database
Feinberg et al. (2018)	GNN	Predict protein–ligand binding affinity	AUC = 85.7%	Open databases: QM8 and GDB-8
Wang et al. (2022)	GNN	Predict several molecular properties	AUC = 92.8%	Unknown
Lane et al. (2020)	RF, k-NN, SVM, NB, Adaboost, DT, RNN	Predict molecular properties	Highest Acc = 84.1%	Open database: ChEMBL
Ashraf et al. (2023)	XGBoost and SHAP	Predict bioactivity	Acc = 93%	Open database: ChEMBL and PubChem
Wallach et al. (2015)	CNN	Predict bioactivity of small molecules	AUC = 90%	Open databases: Directory of Useful Decoys Enhanced (DUDE) benchmark, ChEMBL-20 PMD, etc
Aly and Alotaibi (2023)	RNN	Predict modified gedunin	Acc = 98.68%	Open databases: CHEMBL and Drug Bank
Ahmad et al. (2024)	GNN	Predicting silico solubility	Acc = 0.79%	Open databases: AqSolDB, Lovric and etc

6.2. Pharmacogenomics

Pharmacogenomics studies how genes affect an individual's response to drugs. Recently, supervised ML has emerged as a powerful tool in advancing pharmacogenomics, as depicted in Fig. 8. Ikonnikova et al. (2022) utilized various SL models to explore aspirin resistance (AR) in patients with ischemic stroke. Their research focused on assessing the impact of both clinical and genetic factors on AR.

Athreya et al. (2019) used the RF model in predicting the efficacy of selective serotonin reuptake inhibitors (SSRIs) for treating major depressive disorder (MDD) by incorporating pharmacogenomic biomarkers alongside clinical measures. Through the combination of genetic information with clinical indicators of depression severity, the study showcases a successful application of SL ML in predicting patient responses to SSRI treatment, potentially guiding more tailored and effective therapeutic strategies for individuals suffering from MDD. Lin et al. (2018) extended the feed-forward neural network as the predictive model for determining antidepressant treatment response in MDD. Their study aimed to differentiate between responders and non-responders to antidepressant treatment and to forecast treatment outcomes by analyzing a complex array of genetic and clinical data. This array included single nucleotide polymorphisms (SNPs), alongside demographic and clinical indicators such as age, sex, baseline Hamilton Rating Scale for Depression (HRSD) scores, history of depressive episodes, marital status, and history of suicide attempts. The work of Pandi et al. (2021) proposed several supervised ML methods for the classification of pharmacogenomic variants, especially focusing on novel and rare variants, by assigning them to specific protein activity predictions, thereby facilitating the prioritization of these variants for potential clinical impact in pharmacogenomics. Tang et al. (2017) compared various SL algorithm-based techniques, including ANN and eight ML methods, in predicting the stable dose of tacrolimus in renal transplant patients, thereby enhancing personalized medicine in

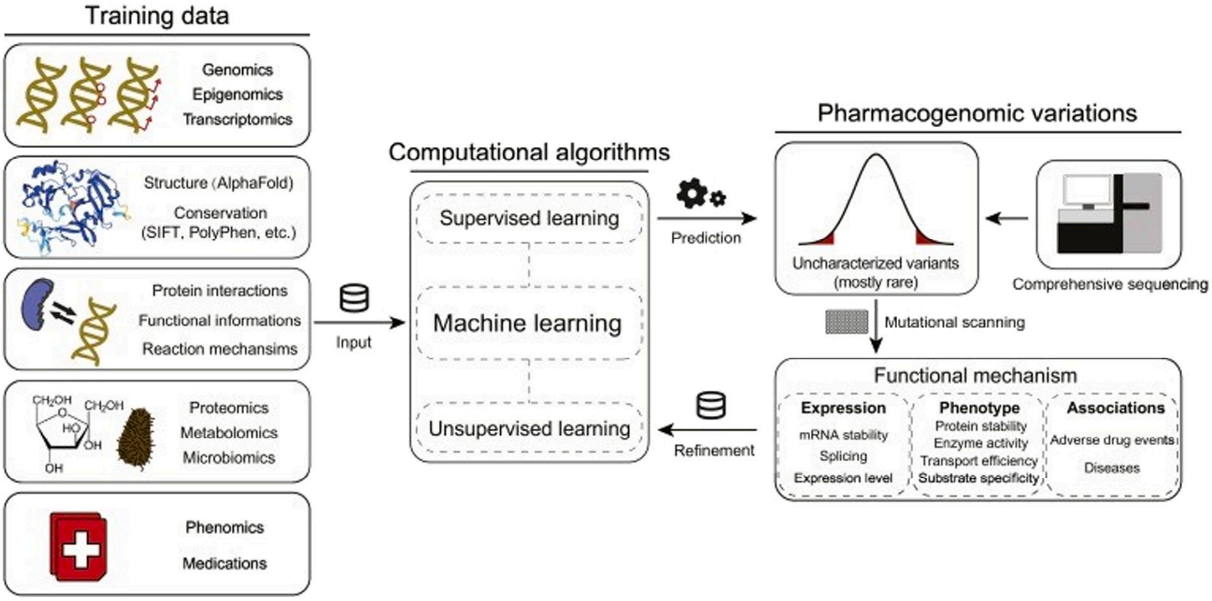

Fig. 8. Machine learning prediction for pharmacogenomics prediction (Zhou, Tremmel, Schaeffeler, Schwab, & Lauschke, 2022).

tacrolimus administration. Thishya, Vattam, Naushad, Raju, and Kutala (2018) applied ANN to predict the bioavailability of tacrolimus and LR to assess the risk for post-transplant diabetes, based on ABCB1 and CYP3A5 genetic polymorphisms. These models aim to explore both individual and synergistic effects of genetic and demographic factors on these outcomes, thereby enhancing personalized treatment plans for transplant recipients. Verhaeghe et al. (2022) developed the Catboost and Gaussian method to predict piperacillin plasma concentrations in critically ill patients with higher accuracy and to provide uncertainty quantification for clinical practice, addressing the limitations of current Population Pharmacokinetic models. The models showed potential improvements in therapeutic drug monitoring and dosing regimen adaptations for piperacillin/tazobactam. In a similar study, Fu et al. (2024) developed several SL models for an individualized dosing model of sertraline for adolescents with depression to account for the variability in pharmacokinetic parameters. The CatBoost model was selected for its superior performance in predicting the optimal sertraline dose based on key variables, thereby offering clinicians a guide to tailor medication regimens more effectively. Lin et al. (2020) used an ensemble approach to predict antidepressant treatment response and remission in patients with major depressive disorder (MDD) by analyzing genetic variants and clinical variables. This ensemble model was designed to differentiate between responders and non-responders to antidepressant treatments, offering a potential bioinformatics tool for personalized treatment planning. For an antidepressant selection study, Sheu et al. (2023) predicted patient responses to several classes of antidepressants including SSRI, SNRI, bupropion, and mirtazapine, using electronic health records and supervised ML techniques, aiming to mitigate the trial-and-error approach traditionally associated with antidepressant selection. By integrating EHR data into their models, the study provides insights into personalized treatment predictions and patient-specific factors affecting antidepressant responses, potentially steering the development of clinical decision support systems for more efficient treatment selection. Table 2 presents a summary of supervised ML algorithms applied in pharmacogenomics.

As indicated in Table 2, common models used for Pharmacogenomics include CatBoost, RF, ANN, AdaBoost, and XGBoost, used in pharmacogenomic studies to predict drug response and efficacy. These models are designed to tailor medical treatments to individual genetic profiles, improving patient outcomes. For example, Ikonnikova et al. used CatBoost and SHAP to develop a predictive model for aspirin resistance, achieving an AUC of 88%, while Athreya et al. (2019) employed RF to predict antidepressant outcomes with AUCs ranging from 70% to 90%.

Key performance indicators include AUC, accuracy, precision, sensitivity, specificity, and correlation. Also, notable performances include Pandi et al. (2021) achieving an accuracy of 85% with AdaBoost, XGBoost, LR, and RF for assessing pharmacogenomic variants, and Fu et al. reporting the highest AUC of 93% with AdaBoost and XGBoost for predicting sertraline response in adolescents. The studies utilize both open and private databases, reflecting the proprietary nature of some pharmacogenomic data and the importance of comprehensive datasets for developing robust predictive models. Methods, such as SHAP enhance the interpretability of predictions, particularly in clinical applications where decisions impact patient health.

6.3. Biomarker identification

Biomarkers encompass a broad range of covariates, such as phenotypical, clinical, gene, and protein expression markers, central to understanding and predicting how different individuals will respond to a given treatment. Many studies have utilized supervised ML techniques to classify patients based on the presence or absence of specific biomarkers, effectively predicting their response to treatment. For instance, Xie et al. (2021) applied various SL algorithms to identify diagnostic biomarkers for lung cancer among Chinese patients, pioneering an interdisciplinary approach that merges metabolomics with ML techniques in lung cancer research. The study demonstrated that ML algorithms support the feasibility of blood-based screening for early lung cancer diagnosis, with the potential to extend this method to other types of cancer. Tabl, Alkhateeb, ElMaraghy, Rueda, and Ngom (2019) used several supervised ML algorithms including RF, SVM, and Naive Bayes, RF for identifying gene biomarkers, guiding the treatment of breast cancer. Through classification and feature selection techniques, the goal was to train a model that could effectively predict sample labels based on gene expression profiles. Ghosh et al. (2022) applied the RF model, together with other models to identify the most significant metabolomic biomarkers in blood for lung cancer prediction, utilizing Plasma and Serum samples through a two-phase process involving initial tests for potential biomarker determination and Recursive Feature Elimination with RF for final biomarker identification. Hajjo et al. (2021) used several SL models to identify disease-specific MRI biomarkers for cancer diagnosis,

Table 2
Summary of supervised machine learning algorithms used for Pharmacogenomics.

Reference	Model	Scope	Performance	Database
Ikonnikova et al. (2022)	CatBoost and SHAP	Develop a predictive model for aspirin resistance	AUC = 88%	Unknown
Athreya et al. (2019)	RF	Predict antidepressant outcomes.	AUC = 70%–90%, 75%–90%	Open database: PGRN-AMPS, STAR*D and ISPC
Lin et al. (2018)	ANN	Predict antidepressant response in major depression.	AUC = 82%, Sen = 75%, Spec = 69%	Private
Pandi et al. (2021)	AdaBoost, XGBoost, LR, RF	Assess pharmacogenomic variant.	Acc = 85%, Precision = 85%, Sen = 84%, Spec = 94%	Open database: PharmGKB
Tang et al. (2017)	ANN, RF, etc	Predict tacrolimus dose in renal transplant recipients.	Highest Corr (R) = 73%	Private
Thishya et al. (2018)	ANN and LR	Predict bioavailability of tacrolimus in patients	Corr (R) = 93%–96%	Private
Verhaeghe et al. (2022)	CatBoost and Gaussian method	Predict piperacillin plasma concentrations.	Corr (R) = 31.94–0.64 and 33.53–0.60	Private
Fu et al. (2024)	AdaBoost, XGBoost, DT, etc	Predict sertraline in adolescents.	Highest AUC = 93%	Private
Lin et al. (2020)	LR, SVM, DT, etc	Predict antidepressant treatment response.	Highest AUC = 81%	Private
Sheu et al. (2023)	RF, GBM, etc with SHAP	Predict differential response to antidepressant classes.	Highest AUC = 70%	Private

prognosis prediction, and treatment efficacy assessment, addressing the need for reliable and non-invasive oncology biomarkers with high specificity. Another study by Abeel et al. (2010) used the ensemble feature selection techniques within SVMs, aimed at increasing the stability of selected biomarkers while enhancing classification performance, which is crucial for reliable diagnosis and prognosis models in biomedical applications. A similar study by Gutiérrez-Gómez et al. (2020) utilized the recursive feature elimination method, in conjunction with SVM, to identify stable biomarkers for schizophrenia across structural, functional, and multi-modal connectomes of both healthy controls and patients, ensuring accuracy and stability across various dataset subsamplings. Zhang and Liu (2019) utilized recursive feature reduction and RF classification to identify biomarker genes across 12 types of cancers. The study analyzed classification effects in control and disease samples using high-throughput '-omic' datasets, such as RNA-sequencing data from the Cancer Genome Atlas (TCGA). The models were used to select a parsimonious set of genes with the highest classification accuracy, ultimately evaluated through tenfold cross-validation. This process revealed insights into the dysfunctional and pathogenic mechanisms associated with the identified biomarkers. Salvatore et al. (2015) applied SVM and feature selection methods to identify magnetic resonance (MR)-related biomarkers for the in vivo differential diagnosis of Alzheimer's Disease (AD). The study specifically focused on distinguishing between patients with AD, those with mild cognitive impairment (MCI) who will or will not convert to AD, and healthy controls. The optimized ML algorithm utilized morphological T1-weighted MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, achieving classification accuracies by identifying critical brain regions involved in AD's pathophysiological mechanisms. This supports the application of computer-based diagnosis in the early management of AD patients. Rehman et al. (2019) used the SVM and RF models together with feature selection methods for validating and ranking the importance of certain small noncoding microRNAs (miRNAs) as biomarkers for breast cancer. The study concluded that ML plays a pivotal role in detecting and diagnosing cancer through the analysis of miRNAs. Chen et al. (2014) applied several supervised ML algorithms including SVM, DT and other models for the identification Gene selection for cancer identification. Their approach focused on selecting a small, informative subset of genes from thousands of candidates using computational intelligence methods to analyze microarray data, significantly contributing to the understanding of gene involvement in cancer occurrence.

Table 3 presents a summary of supervised ML algorithms applied in biomarker identification. Overall, Table 3 highlights the successful application of various supervised ML algorithms for identifying biomarkers across multiple diseases. The high performance demonstrates the models' potential to make reliable predictions. The reliance on diverse data sources ensures comprehensive and robust model training, while the integration of feature selection methods enhances model interpretability and the identification of key biomarkers, which is essential for advancing personalized medicine and improving patient outcomes.

6.4. Pharmacokinetics and pharmacodynamics (PK/pd) modeling

The application of supervised ML in Pharmacokinetics and Pharmacodynamics (PK/PD) Modeling has shown promising results in recent studies (Ahmadi, Alizadeh, Ayyoubzadeh, & Abiyarghamsari, 2024; Al-Bahou, Bruner, Moore, & Zarrinpar, 2024; Degraeve et al., 2024; Ponthier et al., 2023; Sánchez-Herrero, Calvet, & Juan, 2023). Yang, Smith, Patel, and Lee (2019) applied ML methods, such as RF and SVM, to predict the pharmacokinetics of tacrolimus in kidney transplant patients. Their models exhibited strong predictive performance, indicating the potential of ML in improving drug dosing strategies. Liu, Wang, Zhang, and Chen (2020) provided a comprehensive review of ML in drug discovery and development, including PK/PD modeling. The study highlighted the significant impact of ML across various stages of drug development, emphasizing its role in accelerating the drug discovery process. Nguyen, Tran, Le, and Phan (2021) introduced a deep reinforcement learning approach for personalized dosing of tacrolimus in kidney transplant patients. Their model learned optimal dosing strategies based on patient-specific characteristics, demonstrating the potential of deep learning in personalized medicine. Gupta, Sharma, Kumar, and Singh (2019) the integration of ML techniques with mechanistic modeling approaches in PK/PD modeling was discussed. They emphasized the synergistic effects of combining these methods, leading to improved model predictions and decision-making in drug development. Similarly, Napolitano, Rossi, Bianchi, and Esposito (2018) reviewed computational tools and approaches for predictive modeling of drug response, including ML. They highlighted the application of these methods in PK/PD modeling and personalized medicine, underscoring their importance in advancing pharmacological research. Overall, these works underscore the growing importance and potential

Table 3
Summary of supervised machine learning algorithms used for Biomarker Identification.

Reference	Model	Scope	Performance	Database
Xie et al. (2021)	NB	Diagnostic biomarkers for lung cancer among Chinese patients	AUC = 98%, Sen = 98.1%, Spec = 100%	Metabolomics in lung cancer research
Zhang et al. (2021)	RF, SVM, and NB, etc	Identifying genes markers for breast cancer treatment	Highest Acc = 100%	Combined databases
Ghosh et al. (2022)	RF, XGBoost, other models	Metabolomic biomarkers for lung cancer	Acc = 100% and 91%	Public database
Bhat (2017)	SVM	Bioinformatic gene expression analysis	Acc = 97%	Unknown
Abeel et al. (2010)	SVM, ensemble feature selection techniques	Biomarkers for biomedical applications	Acc = 100%	Unknown
Gutiérrez-Gómez et al. (2020)	SVM, recursive feature elimination	Biomarkers for schizophrenia	Acc = 79%	Structural, functional, multi-modal connectomes
Zhang and Liu (2019)	RF, recursive feature reduction	Biomarker genes across 12 types of cancers	Highest AUC = 99.6%	High-throughput omic data, TCGA
Salvatore et al. (2015)	SVM, feature selection methods	MR-related biomarkers for Alzheimer's Disease	Highest Acc = 76%	T1-weighted MRI data, ADNI cohort
Rehman et al. (2019)	SVM, RF, feature selection methods	miRNAs as biomarkers for breast cancer	Highest AUC = 99.9%	Private
Chen et al. (2014)	Naive Bayes, DT, SVM, etc	Gene selection for cancer identification:	Highest Acc = 97.68	Public data: Microarray cancer datasets

of ML in advancing PK/PD modeling and personalized medicine. By leveraging ML techniques, researchers can enhance drug dosing strategies, improve model predictions, and accelerate the drug development process, ultimately leading to better healthcare outcomes.

6.5. *Chemical synthesis prediction and retrosynthesis planning*

The application of supervised ML in chemical synthesis prediction and retrosynthesis planning has gained significant attention in recent years, offering promising advancements in drug discovery and materials science (Ali, Meng, Khan, & Jiang, 2024; Ding et al., 2024; Zhong et al., 2024). Various works have explored the use of ML models to predict chemical reactions, propose novel synthesis pathways, and optimize reaction conditions (Griffin, Coley, Frank, Hawkins, & Jensen, 2023; Mikolajczyk et al., 2023; Yu et al., 2023). One notable study by Coley, Barzilay, Jaakkola, Green, and Jensen (2019) demonstrated the effectiveness of GNNs in predicting chemical reaction outcomes. By encoding molecular structures as graphs, GNNs can capture the underlying chemical relationships and predict reactions with high accuracy. This approach has shown great potential in accelerating the discovery of new chemical compounds and optimizing synthetic routes. Another recent work by Schwaller, Gaudin, Lanyi, Bekas, and Laino (2020) introduced a deep learning model, called ReLeaSE (Retrosynthetic Library of Synthetic Exemplars), for retrosynthesis planning. ReLeaSE utilizes a generative adversarial network (GAN) to propose synthetic routes for target molecules based on a library of known reactions. The model's ability to generate diverse and synthetically feasible pathways has demonstrated its utility in automated synthesis planning. In addition to reaction prediction and retrosynthesis planning, supervised ML has been applied to optimize reaction conditions and predict reaction outcomes under specific constraints. For example, a study by Wei, Hu, Yang, and Lei (2021) utilized a Bayesian optimization approach to optimize reaction conditions for palladium-catalyzed cross-coupling reactions. The model successfully identified optimal conditions that improved reaction efficiency and selectivity. These works highlight the potential of supervised ML in advancing chemical synthesis prediction and retrosynthesis planning. By leveraging large datasets of chemical reactions and molecular structures, ML models can provide valuable insights and guidance for accelerating the discovery and development of new chemical compounds.

7. Notable challenges in supervised learning for drug discovery and development

Despite the success of SL in various applications, there are notable challenges that researchers and practitioners encounter when working with SL algorithms in drug discovery and development. This section discusses some of these challenges and potential solutions.

7.1. *Overfitting*

Overfitting is a significant challenge in SL. It occurs when an ML model achieves excellent performance on the training instances but performs poorly on unseen instances (Mienye & Sun, 2021). Overfitting arises due to the model's excessive complexity, which makes it memorize the training set instead of learning the hidden relationships. In drug discovery, this can lead to models that perform well on known compounds but fail to generalize to novel compounds. One commonly used approach to mitigate this problem is regularization, which involves adding a penalty term to the objective function of the model. A popular regularization technique is L1 regularization, which adds a penalty term proportional to the sum of the absolute values of the model's coefficients. This technique can help select the most relevant features for predicting drug efficacy and safety, improving the model's generalizability (Ma, Miao, Niu, & Zhang, 2019).

7.2. *Imbalanced data*

Imbalanced data refers to a scenario where the class distributions in the training data are skewed (He & Garcia, 2009). In drug discovery, this is common when dealing with datasets where positive instances (e.g., active compounds) are much less frequent than negative instances (e.g., inactive compounds). This imbalance can bias the model towards the majority class, leading to poor performance in identifying potentially active compounds. Techniques such as data resampling, cost-sensitive learning, and ensemble learning have been developed to address this issue. Data resampling techniques, like SMOTE and ADASYN, can balance class distributions by replicating minority class samples (Wang, Dai, Shen, & Xuan, 2021; Zakariah, AlQahtani, & Al-Rakhami, 2023). Cost-sensitive learning assigns different misclassification costs to each class to prioritize the correct classification of minority class instances (Mienye et al., 2021). Ensemble methods, such as bagging and boosting, combine multiple models to improve performance on imbalanced datasets (Obaido et al., 2024).

7.3. Bias and fairness

SL algorithms are trained on historical data that may contain biases, reflecting societal inequalities and prejudices (Okolo, Aruleba, & Obaido, 2023). In drug discovery, biased data can lead to models that perform better for certain populations and worse for others, raising ethical concerns. Addressing bias and ensuring fairness in drug discovery models is crucial for equitable healthcare outcomes. Techniques such as data augmentation and balancing can help mitigate biases by ensuring diverse and representative input data (Connor, Khoshgoftaar, & Borko, 2021; Dao et al., 2019; Rebuffi et al., 2021; Summers & Dinneen, 2019). Algorithmic modifications, like fairness-aware learning, incorporate fairness constraints into the learning process to promote fair decision-making (Caton & Haas, 2020; Mary, Calauzenes, & El Karoui, 2019; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). Post-processing methods, such as calibration, adjust model predictions to meet fairness criteria, providing more equitable outcomes.

7.4. Interpretability

While SL algorithms are powerful in making accurate predictions, they often lack transparency in how predictions are made (Hong, Hullman, & Bertini, 2020). In drug discovery, interpretability is essential for gaining insights into why a model predicts certain compounds as effective or safe. This understanding can build trust in the model, help identify potential biases, and meet regulatory requirements. Techniques such as feature importance, permutation importance, partial dependence plots, and SHAP values can enhance interpretability by identifying the most significant features in the model's decision-making process (Carletti, Terzi, & Susto, 2023; Kang, Koo, & Ryu, 2022; Moreno-Sanchez, 2020; Obaido, Ogbuokiri, Mienye et al., 2022).

8. Discussions and future research directions

Supervised learning algorithms have significantly impacted drug discovery and development by enabling predictions and decisions based on labeled training data.

The summarized tables in this study highlight several critical insights into the application of supervised learning in drug discovery and development. Firstly, a diverse range of supervised learning models, including RF, SVM, and GNN, have been effectively utilized across various aspects of drug discovery, such as predicting molecular properties, protein–ligand binding affinity, and bioactivity. This variety showcases the versatility and adaptability of these algorithms to different types of biomedical data and prediction tasks.

Performance metrics such as accuracy and AUC were commonly used to evaluate these models, with several studies achieving high accuracy (e.g., 98.68% with RNN for predicting modified gedunin) and strong AUC scores (e.g., 92.8% with GNN for predicting molecular properties). These metrics provide a clear indication of the models' effectiveness in making accurate predictions, which is crucial for advancing drug discovery processes and improving therapeutic outcomes. Additionally, the availability and quality of data play a significant role in the performance of these models. The studies leveraged a mix of open and private databases, including ChEMBL, PubChem, and the Protein Data Bank, Indicating the importance of comprehensive and high-quality datasets in training robust and generalizable models. Meanwhile, techniques such as SHAP were employed to enhance model interpretability, which is critical for understanding the predictions made by complex models and ensuring their reliability in practical applications.

Despite the advancements recorded in the literature, new challenges arise as technology advances. Therefore, future research can explore diverse areas. Firstly, future research can focus on developing more efficient algorithms and techniques to address bias and fairness in drug discovery (Mehrabi et al., 2021). Ensuring fairness and preventing discrimination requires exploring methods to mitigate bias, such as algorithmic fairness techniques and data preprocessing methods. Developing transparent and interpretable models is also crucial for understanding and mitigating bias in supervised learning algorithms. Secondly, transfer learning techniques hold promise for enhancing supervised learning algorithms' performance and generalization capabilities, especially in scenarios where obtaining labeled data is expensive or time-consuming. Future research could focus on developing robust transfer learning techniques that can effectively transfer knowledge across domains and adapt to different data distributions while minimizing domain shift.

Additionally, developing more robust and scalable algorithms to handle large-scale datasets is essential. With the increasing availability of big data in drug discovery, traditional supervised learning algorithms may struggle to process and learn from vast amounts of data effectively. Future research can focus on developing algorithms that efficiently handle large-scale datasets, such as distributed learning algorithms or algorithms leveraging parallel computing architectures. Designing algorithms that can adapt and learn in real time from streaming data is also important, as streaming data, including sensor data and social media feeds, have become prevalent in various domains.

Lastly, integrating supervised learning algorithms with other ML approaches, such as unsupervised learning or reinforcement learning, could be beneficial. Combining different learning paradigms can lead to more powerful models that can deal with various tasks and data types. For example, unsupervised learning algorithms can pre-train a model on unlabeled data, which can then be fine-tuned using labeled data through SL. Similarly, integrating reinforcement learning with supervised learning can enable the development of intelligent systems that can make predictions based on labeled data and learn from feedback to optimize their decision-making process.

9. Conclusion

Supervised learning has been widely adopted in the field of ML, allowing for the development of robust models capable of making accurate predictions and decisions. This paper presents a concise and comprehensive overview of supervised learning in the drug design and development field, including the widely used algorithms, challenges, and future research directions. It covers key categories of supervised learning algorithms such as probabilistic classifiers, linear classifiers, deep learning, and boosting algorithms applied in this field. This paper will be beneficial to researchers, practitioners, and students who are interested in understanding and applying ML techniques to enhance drug discovery and development processes. It aims to provide insights into the practical applications of these algorithms and encourage further research in optimizing their effectiveness in the pharmaceutical industry.

CRediT authorship contribution statement

George Obaido: Conceptualization and methodology, Contributing significantly to the writing of the original draft and participating in the review and editing process. **Ibomoiye Domor Mienye:** Conceptualization and methodology, Contributing to the original draft writing, Supervising the project, and engaging in review and editing efforts. **Oluwaseun F. Egbelowo:** Review and editing of the manuscript. **Ikiomoye Douglas Emmanuel:** Investigation and validation of the study. **Adeola Ogunleye:** Formal analysis validation efforts. **Blessing Ogbuokiri:** Resources and project administration, Ensuring the smooth progression of the research. **Pere Mienye:** Resources and project administration, Ensuring the smooth progression of the research. **Kehinde Aruleba:** Resources and project administration, Ensuring the smooth progression of the research.

Table A.4
Acronyms used in the study.

Acronyms	Meaning	Acronyms	Meaning
ML	Machine Learning	AI	Artificial Intelligence
SL	Supervised Learning	RL	Reinforcement Learning
LR	Logistic Regression	DT	Decision Tree
RF	Random Forest	SVM	Support Vector Machine
LASSO	Least Absolute Shrinkage and Selection Operator	NB	Naive Bayes
NN	Neural Network	MLP	Multilayer Perceptron
ANN	Artificial Neural Network	DAG	Directed Acyclic Graph
CART	Classification and Regression Tree	IG	Information Gain
GL	Graph Learning	HL	Hierarchical Learning
GR	Gradient Reversal	HR	Heart Rate
CNN	Convolutional Neural Network	FC	Fully Connected (layer)
RNN	Recurrent Neural Network	LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit	GNN	Graph Neural Network
GNNs	Graph Neural Networks	DL	Deep Learning
TP	True Positive	TN	True Negative
FP	False Positive	FN	False Negative
TPR	True Positive Rate	FPR	False Positive Rate
AUC	Area Under the Curve	ROC	Receiver Operating Characteristic
MAE	Mean Absolute Error	MSE	Mean Squared Error
RMSE	Root Mean Squared Error	COVID	Coronavirus Disease
SARS	Severe Acute Respiratory Syndrome	SHAP	SHapley Additive exPlanations
RGNN	Residual Graph Neural Network	GB	Gradient Boosting
XT	Extra Trees	GDB-8	Generated Database Version 8
DUDE	Decomposing Unobserved Dynamics using Expectations	PMD	Principal Model Decomposition
CHEMBL	Chemical Database	AR	Augmented Reality
MDD	Major Depressive Disorder	SSRI	Selective Serotonin Reuptake Inhibitor
HRSD	Hamilton Rating Scale for Depression	SNRI	Serotonin-Norepinephrine Reuptake Inhibitor
EHR	Electronic Health Record	PGRN	Progranulin
AMPS	Antidepressant Medication Pharmacogenomics Study	STAR	Sequenced Treatment Alternatives to Relieve Depression
ISPC	International SSRI Pharmacogenomics Consortium	GBM	Gradient Boosting Machine
MRI	Magnetic Resonance Imaging	RNA	Ribonucleic Acid
TCGA	The Cancer Genome Atlas	MR	Magnetic Resonance
AD	Alzheimer’s Disease	MCI	Mild Cognitive Impairment
ADNI	Alzheimer’s Disease Neuroimaging Initiative	PK	Pharmacokinetics
PD	Pharmacodynamics	GAN	Generative Adversarial Network
SMOTE	Synthetic Minority Over-sampling Technique	ADASYN	Adaptive Synthetic Sampling Approach
LDA	Linear Discriminant Analysis	PGRN-AMPS	Pharmacogenomics Research Network Antidepressant Medication Pharmacogenomics Study

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

All authors approved the version of the manuscript to be published.

Appendix. Acronyms

See Table A.4.

References

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, *26*(3), 392–398.

Agbele, K. K., Oriogun, P. K., Seluwa, A. G., & Aruleba, K. D. (2015). Towards a model for enhancing ICT4 development and information security in healthcare system. In *2015 IEEE international symposium on technology and society* (pp. 1–6). IEEE.

Ahmad, W., Tayara, H., Shim, H., & Chong, K. T. (2024). SolPredictor: Predicting solubility with residual gated graph neural network. *International Journal of Molecular Sciences*, *25*(2), 715.

Ahmadi, M., Alizadeh, B., Ayyoubzadeh, S. M., & Abiyarghamsari, M. (2024). Predicting pharmacokinetics of drugs using artificial intelligence tools: A systematic review. *European Journal of Drug Metabolism and Pharmacokinetics*, 1–14.

Al-Bahou, R., Bruner, J., Moore, H., & Zarrinpar, A. (2024). Quantitative methods for optimizing patient outcomes in liver transplantation. *Liver Transplantation*, *30*(3), 311–320.

Ali, R. S. A. E., Meng, J., Khan, M. E. I., & Jiang, X. (2024). Machine learning advancements in organic synthesis: A focused exploration of artificial intelligence applications in chemistry. *Artificial Intelligence Chemistry*, *2*(1), Article 100049.

Aly, M., & Alotaibi, A. S. (2023). Molecular property prediction of modified gedunin using machine learning. *Molecules*, *28*(3), 1125.

Anava, O., & Levy, K. (2016). K*-nearest neighbors: From global to local. In *Advances in neural information processing systems*: *vol. 29*.

Arar, Ö. F., & Ayan, K. (2017). A feature dependent naive Bayes approach and its application to the software defect prediction problem. *Applied Soft Computing*, *59*, 197–209.

Aruleba, R. T., Adekiya, T. A., Ayawei, N., Obaido, G., Aruleba, K., Mienye, I. D., et al. (2022). COVID-19 diagnosis: A review of rapid antigen, RT-PCR and artificial intelligence methods. *Bioengineering*, *9*(4), 153.

Aruleba, K., Obaido, G., Ogbuokiri, B., Fadaka, A. O., Klein, A., Adekiya, T. A., et al. (2020). Applications of computational methods in biomedical breast cancer imaging diagnostics: A review. *Journal of Imaging*, *6*(10), 105.

Ashraf, F. B., Akter, S., Mumu, S. H., Islam, M. U., & Uddin, J. (2023). Bio-activity prediction of drug candidate compounds targeting SARS-Cov-2 using machine learning approaches. *Plos one*, *18*(9), Article e0288053.

Athreya, A. P., Neavin, D., Carrillo-Roa, T., Skime, M., Biernacka, J., Frye, M. A., et al. (2019). Pharmacogenomics-driven prediction of antidepressant treatment outcomes: A machine-learning approach with multi-trial replication. *Clinical Pharmacology and Therapeutics*, *106*(4), 855–865.

Austin, P. C., & van Buuren, S. (2023). Logistic regression vs. predictive mean matching for imputing binary covariates. *Statistical Methods in Medical Research*, *32*(11), 2172–2183.

Baştanlar, Y., & Özuysal, M. (2014). Introduction to machine learning. *miRNomics: MicroRNA Biology and Computational Analysis*, 105–128.

Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. *Procedia Computer Science*, *89*, 117–123.

Bentes, C., Velotto, D., & Tings, B. (2017). Ship classification in TerraSAR-X images with convolutional neural networks. *IEEE Journal of Oceanic Engineering*, *43*(1), 258–266.

Bhat, H. F. (2017). Evaluating SVM algorithms for bioinformatic gene expression analysis. *International Journal of Computer Sciences and Engineering*, *6*, 42–52.

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. arXiv preprint arXiv:1809.03006.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Carletti, M., Terzi, M., & Susto, G. A. (2023). Interpretable anomaly detection with diffi: Depth-based feature importance of isolation forest. *Engineering Applications of Artificial Intelligence*, *119*, Article 105730.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *ACM Computing Surveys*.

Chen, D. W., Miao, R., Yang, W. Q., Liang, Y., Chen, H. H., Huang, L., et al. (2019). A feature extraction method based on differential entropy and linear discriminant analysis for emotion recognition. *Sensors*, *19*(7), http://dx.doi.org/10.3390/s19071631, URL https://www.mdpi.com/1424-8220/19/7/1631.

Chen, K. H., Wang, K. J., Tsai, M. L., Wang, K. M., Adrian, A. M., Cheng, W. C., et al. (2014). Gene selection for cancer identification: A decision tree model empowered by particle swarm optimization algorithm. *BMC Bioinformatics*, *15*, 1–10.

Cheong, Q., Au Yeung, M., Quon, S., Concepcion, K., & Kong, J. D. (2021). Predictive modeling of vaccination uptake in US counties: A machine learning–based approach. *Journal of Medical Internet Research*, *23*(11), Article e33231.

Choudhary, R., & Gianey, H. K. (2017). Comprehensive review on supervised machine learning algorithms. In *2017 international conference on machine learning and data science* (pp. 37–43). IEEE.

Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., & Jensen, K. F. (2019). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, *10*(2), 370–377.

Connor, S., Khoshgoftaar, T. M., & Borko, F. (2021). Text data augmentation for deep learning. *Journal of Big Data*, *8*(1).

Crisci, C., Ghattas, B., & Perera, G. (2012). A review of supervised machine learning algorithms and their applications to ecological data. *Ecological Modelling*, *240*, 113–122.

Cuadros-Rodríguez, L., Pérez-Castaño, E., & Ruiz-Samblás, C. (2016). Quality performance metrics in multivariate classification methods for qualitative analysis. *TRAC Trends in Analytical Chemistry*, *80*, 612–624.

Cui, L., Chen, P., Wang, L., Li, J., & Ling, H. (2021). Application of extreme gradient boosting based on grey relation analysis for prediction of compressive strength of concrete. *Advances in Civil Engineering*, *2021*, 1–14.

Dalal, K. R. (2020). Analysing the role of supervised and unsupervised machine learning in iot. In *2020 international conference on electronics and sustainable communication systems* (pp. 75–79). IEEE.

Dao, T., Gu, A., Ratner, A., Smith, V., De Sa, C., & Ré, C. (2019). A kernel theory of modern data augmentation. In *International conference on machine learning* (pp. 1528–1537). PMLR.

Degraeve, A. L., Bindels, L. B., Haufroid, V., Moudio, S., Boland, L., Delongie, K. A., et al. (2024). Tacrolimus pharmacokinetics is associated with gut microbiota diversity in kidney transplant patients: Results from a pilot cross-sectional study. *Clinical Pharmacology and Therapeutics*, *115*(1), 104–115.

Deng, L., & Li, X. (2013). Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(5), 1060–1089.

Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, *24*(3), 596–606.

Dhaliwal, S. S., Nahid, A.-A., & Abbas, R. (2018). Effective intrusion detection system using XGBoost. *Information*, *9*(7), 149.

Ding, Y., Qiang, B., Chen, Q., Liu, Y., Zhang, L., & Liu, Z. (2024). Exploring chemical reaction space with machine learning models: Representation and feature perspective. *Journal of Chemical Information and Modeling*.

Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., et al. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PLoS One*, *11*(2), Article e0148195.

Dongare, A., Kharde, R., Kachare, A. D., et al. (2012). Introduction to artificial neural network. *International Journal of Engineering and Innovative Technology (IJEIT)*, *2*(1), 189–194.

Dybowski, R., & Gant, V. (2001). *Clinical applications of artificial neural networks*: *vol. 200*, (1), Cambridge University Press Cambridge.

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., et al. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials*, *18*(5), 435–441.

El Naqa, I., & Murphy, M. J. (2022). What are machine and deep learning? *Machine and Deep Learning in Oncology, Medical Physics and Radiology*, 3–15.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Fabris, F., Magalhães, J. P. d., & Freitas, A. A. (2017). A review of supervised machine learning applied to ageing research. *Biogerontology*, *18*(2), 171–188.

Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., et al. (2018). PotentialNet for molecular property prediction. *ACS Central Science*, *4*(11), 1520–1530.

Fu, R., Yu, Z., Zhou, C., Zhang, J., Gao, F., Wang, D., et al. (2024). Artificial intelligence-based model for dose prediction of sertraline in adolescents: A real-world study. *Expert Review of Clinical Pharmacology*, *17*(2), 177–187.

Gavankar, S. S., & Sawarkar, S. D. (2017). Eager decision tree. In *2017 2nd international conference for convergence in technology* (pp. 837–840). IEEE.

Gholami, R., & Fakhari, N. (2017). Support vector machine: principles, parameters, and applications. In *Handbook of neural computation* (pp. 515–535). Elsevier.

Ghosh, U. K., Al Abir, F., Rifaat, N., Shovan, S., Sayeed, A., & Hasan, M. A. M. (2022). Most dominant metabolomic biomarkers identification for lung cancer. *Informatics in Medicine Unlocked*, *28*, Article 100824.

Griffin, D. J., Coley, C. W., Frank, S. A., Hawkins, J. M., & Jensen, K. F. (2023). Opportunities for machine learning and artificial intelligence to advance synthetic drug substance process development. *Organic Process Research & Development*, *27*(11), 1868–1879.

Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 1–49.

Gupta, A., Sharma, S., Kumar, R., & Singh, P. (2019). Machine learning in pharmacokinetic and pharmacodynamic modeling: Concepts and applications. *Pharmaceutical Research*, *36*(4), 61–72.

Gutiérrez-Gómez, L., Vohryzek, J., Chiêm, B., Baumann, P. S., Conus, P., Do Cuenod, K., et al. (2020). Stable biomarker identification for predicting schizophrenia in the human connectome. *NeuroImage: Clinical*, *27*, Article 102316.

Hajjo, R., Sabbah, D. A., Bardaweel, S. K., & Tropsha, A. (2021). Identification of tumor-specific MRI biomarkers using machine learning (ML). *Diagnostics*, *11*(5), 742.

Haldorai, A., Ramu, A., & Suriya, M. (2020). Organization internet of things (IoTs): Supervised, unsupervised, and reinforcement learning. *Business Intelligence for Enterprise Internet of Things*, 27–53.

Harris, J. K. (2021). Primer on binary logistic regression. *Family Medicine and Community Health*, *9*(Suppl 1).

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

He, J., Hao, Y., & Wang, X. (2021). An interpretable aid decision-making model for flag state control ship detention based on SMOTE and XGBoost. *Journal of Marine Science and Engineering*, *9*(2), 156.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266.

Hong, S. R., Hullman, J., & Bertini, E. (2020). Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–26.

Hu, L., Chen, J., Vaughan, J., Aramideh, S., Yang, H., Wang, K., et al. (2021). Supervised machine learning techniques: An overview with applications to banking. *International Statistical Review*, *89*(3), 573–604.

Huang, M. W., Chen, C. W., Lin, W. C., Ke, S. W., & Tsai, C. F. (2017). SVM and SVM ensembles in breast cancer prediction. *PLoS One*, *12*(1), Article e0161501.

Ikonnikova, A., Anisimova, A., Galkin, S., Gunchenko, A., Abdukhalikova, Z., Filippova, M., et al. (2022). Genetic association study and machine learning to investigate differences in platelet reactivity in patients with acute ischemic stroke treated with aspirin. *Biomedicines*, *10*(10), 2564.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31–44.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Unsupervised learning. In *An introduction to statistical learning: with applications in python* (pp. 503–556). Springer.

Janakiraman, A. K., Khanna, K., & Ramkanth, S. (2023). Role of machine learning in automated detection and sorting of pharmaceutical formulations. In *Artificial intelligence in pharmaceutical sciences* (pp. 96–116). CRC Press.

Jantan, I., Ahmad, W., & Bukhari, S. N. A. (2015). Plant-derived immunomodulators: an insight on their preclinical evaluation and clinical trials. *Frontiers in Plant Science*, *6*, Article 158994.

Kang, K. S., Koo, C., & Ryu, H. G. (2022). An interpretable machine learning approach for evaluating the feature importance affecting lost workdays at construction sites. *Journal of Building Engineering*, *53*, Article 104534.

Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021). Review on convolutional neural networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *173*, 24–49.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*: *vol. 30*.

Kelly, A., & Johnson, M. A. (2021). Investigating the statistical assumptions of Naïve Bayes classifiers. In *2021 55th annual conference on information sciences and systems* (pp. 1–6). IEEE.

Koivisto, A. P., Belvisi, M. G., Gaudet, R., & Szallasi, A. (2022). Advances in TRP channel drug discovery: from target validation to clinical studies. *Nature reviews Drug discovery*, *21*(1), 41–59.

Koyejo, O. O., Natarajan, N., Ravikumar, P. K., & Dhillon, I. S. (2014). Consistent binary classification with generalized performance metrics. In *Advances in neural information processing systems*: In *Advances in neural information processing systems.vol. 27*, Kukreja, H., Bharath, N., Siddesh, C., & Kuldeep, S. (2016). An introduction to artificial neural network. *International Journal of Advance Research and Innovative Ideas in Education*, *1*, 27–30.

Kumar, Y., Kaur, K., & Singh, G. (2020). Machine learning aspects and its applications towards different research areas. In *2020 international conference on computation, automation and knowledge management* (pp. 150–156). IEEE.

Kyrimi, E., McLachlan, S., Dube, K., Neves, M. R., Fahmi, A., & Fenton, N. (2021). A comprehensive scoping review of Bayesian networks in healthcare: Past, present and future. *Artificial Intelligence in Medicine*, *117*, Article 102108.

Lane, T. R., Foil, D. H., Minerali, E., Urbina, F., Zorn, K. M., & Ekins, S. (2020). Bioactivity comparison across multiple machine learning algorithms using over 5000 datasets for drug discovery. *Molecular Pharmaceutics*, *18*(1), 403–415.

Li, Y., & Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, *8*(10), 1756.

Lin, E., Kuo, P. H., Liu, Y. L., Yu, Y. W. Y., Yang, A. C., & Tsai, S. J. (2018). A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Frontiers in Psychiatry*, *9*, Article 367995.

Lin, E., Kuo, P. H., Liu, Y. L., Yu, Y. W. Y., Yang, A. C., & Tsai, S. J. (2020). Prediction of antidepressant treatment response and remission using an ensemble machine learning framework. *Pharmaceuticals*, *13*(10), 305.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11–26.

Liu, L., Wang, H., Zhang, Q., & Chen, X. (2020). Machine learning in drug discovery and development: A review. *Drug Development Research*, *81*(8), 874–882.

Ma, R., Miao, J., Niu, L., & Zhang, P. (2019). Transformed L1 regularization for learning sparse deep neural networks. *Neural Networks*, *119*, 286–298.

Mak, K. K., Wong, Y. H., & Pichika, M. R. (2023). Artificial intelligence in drug discovery and development. In *Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays* (pp. 1–38). Springer.

Marchetti, F., Moroni, E., Pandini, A., & Colombo, G. (2021). Machine learning prediction of allosteric drug activity from molecular dynamics. *The Journal of Physical Chemistry Letters*, *12*(15), 3724–3732.

Mary, J., Calauzenes, C., & El Karoui, N. (2019). Fairness-aware learning for continuous attributes and treatments. In *International conference on machine learning* (pp. 4382–4391). PMLR.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Mienye, I. D., & Jere, N. (2024). A survey of decision trees: Concepts, algorithms, and applications. *IEEE Access*.

Mienye, I. D., Obaido, G., Aruleba, K., & Dada, O. A. (2021). Enhanced prediction of chronic kidney disease using feature selection and boosted classifiers. In *International conference on intelligent systems design and applications* (pp. 527–537). Springer.

Mienye, I. D., & Sun, Y. (2021). Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked*, *25*, Article 100690.

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129–99149.

Mienye, I. D., & Sun, Y. (2023). A deep learning ensemble with data resampling for credit card fraud detection. *IEEE Access*, *11*, 30628–30638.

Mienye, I. D., Sun, Y., & Wang, Z. (2019). Prediction performance of improved decision tree-based algorithms: A review. *Procedia Manufacturing*, *35*, 698–703.

Mikolajczyk, A., Zhdan, U., Antoniotti, S., Smolinski, A., Jagiello, K., Skurski, P., et al. (2023). Retrosynthesis from transforms to predictive sustainable chemistry and nanotechnology: A brief tutorial review. *Green Chemistry*, *25*(8), 2971–2991.

Moreno-Sanchez, P. A. (2020). Features importance to improve interpretability of chronic kidney disease early diagnosis. In *2020 IEEE international conference on big data (big data)* (pp. 3786–3792). IEEE.

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference* (pp. 15–25). Springer.

Napolitano, A., Rossi, M., Bianchi, G., & Esposito, M. (2018). Predictive modeling of drug response: A review of computational tools and approaches. *Frontiers in Pharmacology*, *9*, 1–12.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51–62.

Nguyen, N., Tran, M., Le, T., & Phan, K. (2021). Deep reinforcement learning for personalized dosing: An application to tacrolimus dosing in kidney transplant patients. *Journal of Pharmacological Sciences*, *139*(2), 112–120.

Obaido, G., Ogbuokiri, B., Chukwu, C. W., Osaye, F. J., Egbelowo, O. F., Uzochukwu, M. I., et al. (2024). An improved ensemble method for predicting hyperchloremia in adults with diabetic ketoacidosis. *IEEE Access*.

Obaido, G., Ogbuokiri, B., Mienye, I. D., & Kasongo, S. M. (2022). A voting classifier for mortality prediction post-thoracic surgery. In *International conference on intelligent systems design and applications* (pp. 263–272). Springer.

Obaido, G., Ogbuokiri, B., Swart, T. G., Ayawei, N., Kasongo, S. M., Aruleba, K., et al. (2022). An interpretable machine learning approach for hepatitis b diagnosis. *Applied Sciences*, *12*(21), 11127.

Okolo, C. T., Aruleba, K., & Obaido, G. (2023). *Responsible AI in Africa: Challenges and opportunities* (pp. 35–64). Springer International Publishing Cham.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., Akinjobi, J., et al. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, *48*(3), 128–138.

Pandi, M. T., Koromina, M., Tsafaridis, I., Patsilinakos, S., Christoforou, E., van der Spek, P. J., et al. (2021). A novel machine learning-based approach for the computational functional assessment of pharmacogenomic variants. *Human Genomics*, *15*, 1–13.

Patel, K., & Patel, H. B. (2021). A comparative analysis of supervised machine learning algorithm for agriculture crop prediction. In *2021 fourth international conference on electrical, computer and communication technologies* (pp. 1–5). IEEE.

Ponthier, L., Marquet, P., Moes, D. J. A., Rostaing, L., van Hoek, B., Monchaud, C., et al. (2023). Application of machine learning to predict tacrolimus exposure in liver and kidney transplant patients given the MeltDose formulation. *European Journal of Clinical Pharmacology*, *79*(2), 311–319.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In *Advances in neural information processing systems*: *vol. 31*.

Rebuffi, S. A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, *34*, 29935–29948.

Rehman, O., Zhuang, H., Muhamed Ali, A., Ibrahim, A., & Li, Z. (2019). Validation of miRNAs as breast cancer biomarkers with a machine learning approach. *Cancers*, *11*(3), 431.

Rothman, D. (2018). *Artificial intelligence by example: develop machine intelligence from scratch using real artificial intelligence use cases*. Packt Publishing Ltd.

Rubin, A. E., Tummala, S., Both, D. A., Wang, C., & Delaney, E. J. (2006). Emerging technologies supporting chemical process R&D and their increasing impact on productivity in the pharmaceutical industry. *Chemical Reviews*, *106*(7), 2794–2810.

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., et al. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access*, *8*, 101489–101499.

Salvatore, C., Cerasa, A., Battista, P., Gilardi, M. C., Quattrone, A., & Castiglioni, I. (2015). Magnetic resonance imaging biomarkers for the early diagnosis of alzheimer's disease: A machine learning approach. *Frontiers in Neuroscience*, *9*, Article 144798.

Sánchez-Herrero, S., Calvet, L., & Juan, A. A. (2023). Machine learning models for predicting personalized tacrolimus stable dosages in pediatric renal transplant patients. *BioMedInformatics*, *3*(4), 926–947.

Sarkar, C., Das, B., Rawat, V. S., Wahlang, J. B., Nongpiur, A., Tiewsoh, I., et al. (2023). Artificial intelligence and machine learning technology driven modern drug discovery and development. *International Journal of Molecular Sciences*, *24*(3), 2026.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C., & Laino, T. (2020). Molecular transformer for chemical reaction prediction and uncertainty estimation. *Chemical Science*, *11*(13), 3316–3325.

Seixas, F. L., Zadrozny, B., Laks, J., Conci, A., & Saade, D. C. M. (2014). A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. *Computers in Biology and Medicine*, *51*, 140–158.

Selekman, J. A., Qiu, J., Tran, K., Stevens, J., Rosso, V., Simmons, E., et al. (2017). High-throughput automation in chemical process development. *Annual Review of Chemical and Biomolecular Engineering*, *8*, 525–547.

Seng, J. K. P., & Ang, K. L. M. (2017). Big feature data analytics: Split and combine linear discriminant analysis (SC-LDA) for integration towards decision making analytics. *IEEE Access*, *5*, 14056–14065. http://dx.doi.org/10.1109/ACCESS.2017.2726543.

Sevinç, E. (2022). An empowered AdaBoost algorithm implementation: A COVID-19 dataset study. *Computers & Industrial Engineering*, *165*, Article 107912.

Sheu, Y. h., Magdamo, C., Miller, M., Das, S., Blacker, D., & Smoller, J. W. (2023). AI-assisted prediction of differential response to antidepressant classes using electronic health records. *NPJ Digital Medicine*, *6*(1), 73.

Shinde, P. P., & Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 fourth international conference on computing communication control and automation* (pp. 1–6). IEEE.

Sindhu Meena, K., & Suriya, S. (2020). A survey on supervised and unsupervised learning techniques. In *Proceedings of international conference on artificial intelligence, smart grid and smart city applications: AISGSC 2019* (pp. 627–644). Springer.

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between multinomial and Bernoulli naïve Bayes for text classification. In *2019 international conference on automation, computational and technology management* (pp. 593–596). IEEE.

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development* (pp. 1310–1315). Ieee.

Summers, C., & Dinneen, M. J. (2019). Improved mixed-example data augmentation. In *2019 IEEE winter conference on applications of computer vision* (pp. 1262–1270). IEEE.

Suthaharan, S., & Suthaharan, S. (2016). Support vector machine. In *Machine learning models and algorithms for big data classification: thinking with examples for effective learning* (pp. 207–235). Springer.

Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Frontiers in Genetics*, *10*, 256.

Tang, J., Liu, R., Zhang, Y. L., Liu, M. Z., Hu, Y. F., Shao, M. J., et al. (2017). Application of machine-learning models to predict tacrolimus stable dose in renal transplant recipients. *Scientific Reports*, *7*(1), 42192.

Tayyebi, A., Alshami, A. S., Rabiei, Z., Yu, X., Ismail, N., Talukder, M. J., et al. (2023). Prediction of organic compound aqueous solubility using machine learning: A comparison study of descriptor-based and fingerprints-based models. *Journal of Cheminformatics*, *15*(1), 99.

Thishya, K., Vattam, K. K., Naushad, S. M., Raju, S. B., & Kutala, V. K. (2018). Artificial neural network model for predicting the bioavailability of tacrolimus in patients with renal transplantation. *PLoS One*, *13*(4), Article e0191921.

Thomaz, C. E., Kitani, E. C., & Gillies, D. F. (2006). A maximum uncertainty LDA-based approach for limited sample size problems — with application to face recognition. *Journal of the Brazilian Computer Society*, *12*(2), 7–18. http://dx.doi.org/10.1007/bf03192391.

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, *18*(6), 463–477.

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., et al. (2019). Sample size for binary logistic prediction models: beyond events per variable criteria. *Statistical Methods in Medical Research*, *28*(8), 2455–2474.

Verhaeghe, J., Dhaese, S. A., De Corte, T., Vander Mijnsbrugge, D., Aardema, H., Zijlstra, J. G., et al. (2022). Development and evaluation of uncertainty quantifying machine learning models to predict piperacillin plasma concentrations in critically ill patients. *BMC Medical Informatics and Decision Making*, *22*(1), 224.

Wallach, I., Dzamba, M., & Heifets, A. (2015). AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855.

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *11*(1), 24039.

Wang, Z., Liu, M., Luo, Y., Xu, Z., Xie, Y., Wang, L., et al. (2022). Advanced graph and sequence neural networks for molecular property prediction and drug discovery. *Bioinformatics*, *38*(9), 2579–2586.

Wei, J., Hu, Z., Yang, Z., & Lei, A. (2021). Machine learning in chemical reaction optimization. *Chemical Science*, *12*(5), 1686–1695.

Wu, S., Sun, F., Zhang, W., Xie, X., & Cui, B. (2022). Graph neural networks in recommender systems: A survey. *ACM Computing Surveys*, *55*(5), 1–37.

Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., et al. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational Oncology*, *14*(1), Article 100907.

Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48–59.

Xu, L., Pan, S., Xia, L., & Li, Z. (2023). Molecular property prediction by combining LSTM and GAT. *Biomolecules*, *13*(3), 503.

Yang, Y., Smith, J., Patel, R., & Lee, S. (2019). Application of machine learning methods to predict the pharmacokinetics of tacrolimus in kidney transplant patients. *Journal of Pharmacokinetics and Pharmacodynamics*, *46*(5), 521–532.

Yang, X., Song, Z., King, I., & Xu, Z. (2021). A survey on deep semi-supervised learning. arXiv, arXiv preprint arXiv:2103.00550.

Yildiz, B., Bilbao, J. I., & Sproul, A. B. (2017). A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renewable and Sustainable Energy Reviews*, *73*, 1104–1122.

Yu, T., Boob, A. G., Volk, M. J., Liu, X., Cui, H., & Zhao, H. (2023). Machine learning-enabled retrobiosynthesis of molecules. *Nature Catalysis*, *6*(2), 137–151.

Zakariah, M., AlQahtani, S. A., & Al-Rakhami, M. S. (2023). Machine learning-based adaptive synthetic sampling technique for intrusion detection. *Applied Sciences*, *13*(11), 6504.

Zhang, X., Jonassen, I., & Goksøyr, A. (2021). Machine learning approaches for biomarker discovery using gene expression data. *Bioinformatics*.

Zhang, Z., & Liu, Z. P. (2019). Identifying cancer biomarkers from high-throughput RNA sequencing data by machine learning. In *Intelligent computing theories and application: 15th international conference, ICIC 2019, nanchang, China, August 3–6, 2019, proceedings, part II 15* (pp. 517–528). Springer.

Zhang, Y., Wang, Y., Zhou, W., Fan, Y., Zhao, J., Zhu, L., et al. (2019). A combined drug discovery strategy based on machine learning and molecular docking. *Chemical Biology & Drug Design*, *93*(5), 685–699.

Zheng, H., Xiao, F., Sun, S., & Qin, Y. (2022). Brillouin frequency shift extraction based on AdaBoost algorithm. *Sensors*, *22*(9), 3354.

Zhong, Z., Song, J., Feng, Z., Liu, T., Jia, L., Yao, S., et al. (2024). Recent advances in deep learning for retrosynthesis. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *14*(1), Article e1694.

Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., et al. (2020). Graph neural networks: A review of methods and applications. *AI open*, *1*, 57–81.

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, *10*(5), 593.

Zhou, Y., Tremmel, R., Schaeffeler, E., Schwab, M., & Lauschke, V. M. (2022). Challenges and opportunities associated with rare-variant pharmacogenomics. *Trends in Pharmacological Sciences*, *43*(10), 852–865.

Zhou, Y., Wang, J., Gu, Z., Wang, S., Zhu, W., Acena, J. L., et al. (2016). Next generation of fluorine-containing pharmaceuticals, compounds currently in phase II–III clinical trials of major pharmaceutical companies: new structural trends and therapeutic areas. *Chemical Reviews*, *116*(2), 422–518.

Zou, R. Y., & Schonlau, M. (2018). The random forest algorithm for statistical learning with applications in stata. *The Stata Journal*, *20*(3).