Egyptian Informatics Journal 28 (2024) 100558

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full length article

Deep Learning-Assisted Compound Bioactivity Estimation Framework

Yasmine Eid Mahmoud Yousef [a,*], Ayman El-Kilany [b], Farid Ali [c], Yassin M. Nissan [d,e], Ehab E. Hassanein [b]

[a] Faculty of Computer Science, October University for Modern Sciences and Arts, Cairo, Egypt
[b] Information Systems Department, Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt
[c] Information Technology Department, Faculty of Computers and Artificial Intelligence, Beni-Suef University, Beni-Suef, Egypt
[d] Pharmaceutical chemistry department faculty of pharmacy Cairo University, Cairo, Egypt
[e] Pharmaceutical chemistry department faculty of pharmacy MSA University, Cairo, Egypt

ARTICLE INFO

Keywords:
Drug discovery
Deep learning
Regression
Auto-encoder
Classification

ABSTRACT

Drug Discovery is a highly complicated process. On average, it takes six to twelve years to manufacture a new drug and have the product released in the market. It is of utmost importance to find methods that would accelerate the manufacturing process. This significant challenge in drug development can be addressed using deep learning techniques. The aim of this paper is to propose a deep learning-based framework that can help chemists examine compound biological activity in a more accurate manner. The proposed framework employs autoencoder for data representation of the compounds data, which is then classified using deep neural network followed by building a customized deep regression model to estimate an accurate value of the compound bioactivity. The proposed framework achieved an accuracy of 89% in autoencoder reconstruction error, 79.01% in classification, and MAE of 2.4 while predicting compound bioactivity using deep regression model.

1. Introduction

One of the most important and dramatically changing topics in computer aided drug discovery is the process of utilizing Machine learning. Analyzing and predicting the chemical, biological, and physical properties of new compounds is of great research importance, despite its challenging nature [1].

The drug discovery process represented in Fig. 1 shows the complexity and time consumption to develop a new drug from the original idea to the finished product, as the process requires expensive, time-consuming, and intensive labor cycles of medicinal chemistry synthesis and analysis [2]. The manufacturing of a drug takes from six to twelve years on average to be able to release it in the market; thus, it can be defined as a process with great complexity. The success of any drug cannot be assured after being released despite the massive time, work, and money invested.

Notably, a crucial factor in drug discovery is the value of bioactivity, which refers to the ability of a compound to exert a biological effect [3]. The goal of drug discovery is to identify compounds with specific bioactivities that can be used to treat diseases or improve health outcomes. The process of discovering new drugs with desired bioactivities involves a combination of scientific and technological approaches, including screening large libraries of compounds, identifying potential drug targets, optimizing drug candidates through medicinal chemistry, and conducting preclinical and clinical trials [4].

In general, examinations of bioactivity take place in the early stages of drug discovery to screen compounds for potential therapeutic effects. By then, the examination process usually includes testing the compound's ability to interact with a specific target, such as a protein or enzyme, or measuring compound effects on cellular or organismal systems. Therefore, Chemists work on optimizing the identified compound properties to improve its efficacy and safety. For example, the rate of the drug ingredient absorption known as bioavailability, and the movement of the drug in the body and actions of the body on the drug known as pharmacokinetics [3].

Therefore, effective drug discovery requires a thorough understanding of bioactivity, as well as the underlying biological mechanisms that govern the interaction between drugs and their targets.

Machine learning techniques are considered to be more effective in knowledge extraction from data. They can provide a decent replacement for the extensive use of computational resources than physical models [5]. In drug discovery, machine learning can be primarily applied in order to understand and exploit the relationships between chemical structures and their biological activities or Structure-Activity Relationships (SAR) [6].

* Corresponding author.
E-mail addresses: yyousef@msa.edu.eg (Y.E.M. Yousef), a.elkilany@fci-cu.edu.eg (A. El-Kilany), fared.ali@fcis.bsu.edu.eg (F. Ali), ymohammed@msa.edu.eg (Y.M. Nissan), e.ezat@fci-cu.edu.eg (E.E. Hassanein).

https://doi.org/10.1016/j.eij.2024.100558
Received 16 August 2023; Received in revised form 14 May 2024; Accepted 5 October 2024
Available online 15 October 2024

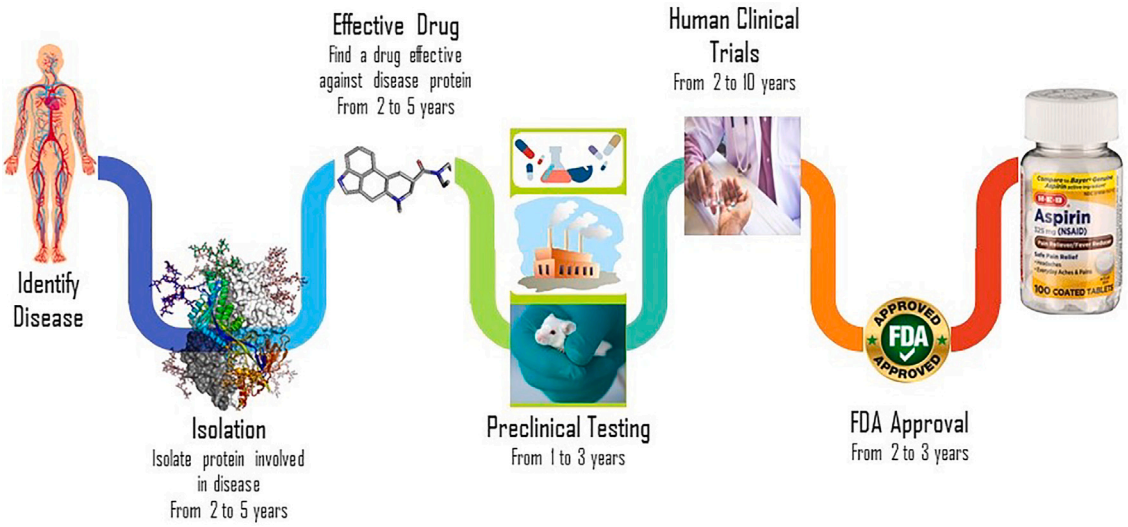


Fig. 1. Drug discovery process.

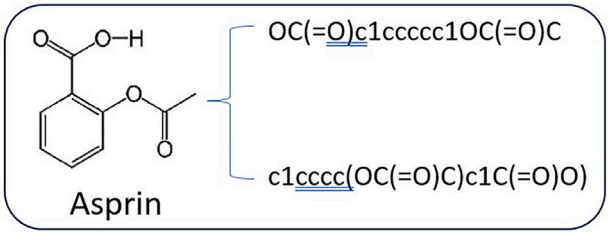


Fig. 2. Asprin different SMILE representation.

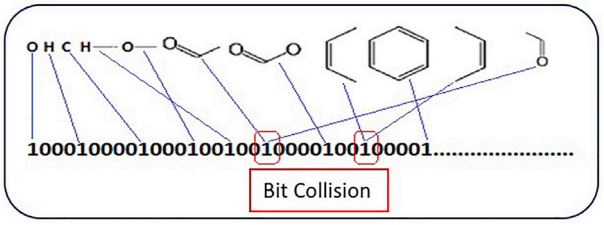


Fig. 3. Asprin hashed Fingerprint (Bit Collision).

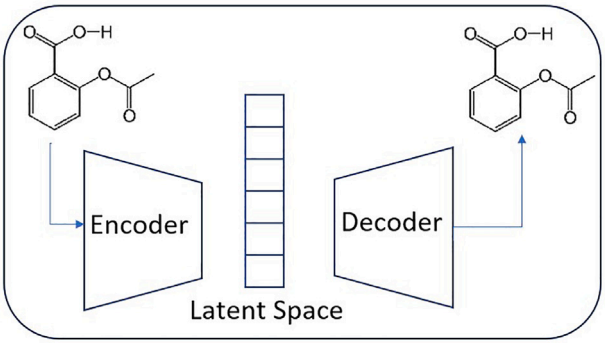


Fig. 4. Asprin representation using autoencoder.

SMILES are amongst the variant ways to represent chemical compound structure in digital manner. There are many different ways to construct the SMILES string for a given molecule. We can start representing the SMILE from different atoms or following a different sequence through the molecule [7]. Hence an Asprin SMILE representation for example can be written in many possible SMILES strings as shown in Fig. 2.

Traditionally, researches use the hashing (finger print) method in order to allow representing the smile in a unique format [8,9]. Fig. 3 shows the hashed fingerprint of the Asprin compound. The generation of this fingerprint representation suffer from long string, bit collision, and time. As SMILES string may be written starting at a different atom or by following a different sequence through the molecule. SMILES have been criticized for not being able to handle relative stereochemistry very well. This paper intends to use deep learning models in order to eventually generate a better representation by which we can estimate bioactivity more accurately. Fig. 4 shows the representation of Asprin compound using autoencoder where the length of its latent space representation reduced with respect to hashing.

Subsequently, section two in this paper presents the related work in which SMILE data representation is covered followed by deep learning models and its utilization in drug discovery. Then, section three outlines the proposed framework including the autoencoder model, deep neural network model, and deep regression model. Thereafter, section four presents the performance evaluation of the proposed framework and experiments details. Finally, the paper ends up by the discussion, conclusion and future work.

2. Related work

Computers and software are used to process, analyze, and interpret chemical data under the domains of chemoinformatics and machine learning, which are interconnected [10]. Data representation is an essential component of chemoinformatics and machine learning, embedded its importance in representing chemical compounds and molecular interactions to be easily processed by computer algorithms [11]. In this study, Chemoinformatics and machine learning are employed in prediction of molecular interactions, discovery and design of novel molecules, and design of drug candidates [12]. In this section, the related work is discussed while focusing on data representation of compounds followed by deep learning models and their utilization in drug discovery.

2.1. Data representation

Data representation has been a profound for many research papers. Gómez-Bombarelli, R., et al. [13] identified a method that change molecules representations from and to multidimensional continuous

representation. The invented model allows the generation of new molecules to efficiently explore and optimize through the chemical compounds open-ended spaces. The researchers observed high fidelity of their autoencoder when constructed SMILES strings and capturing molecular training set characteristic features. In addition, when trained jointly with a property prediction task, the autoencoder showed good power prediction, and the ability of performing molecules gradient-based optimization of the resulted smooth latent space.

Bjerrum, E.J., et al. [14] conducted a pilot study to improve chemical autoencoder latent space and De Novo generation diversity molecu les through the use of hetero-encoders. The researchers used eight atoms fully enumerated train and test set that indicated that the representation of latent space is sensitive to the input and output chosen representations in the training. The better performance is not just limited to SMILES independent representations, but it also provides better description of the chemical space that is relevant to physico-chemical and biological properties.

Songhao Han, et al. introduced a novel approach called Vision-Linguistics coordination Time Sequence-aware News Recommendation (VLSNR) for news recommendation in [15]. The pretrained CLIP (Contrastive Language–Image Pretraining) encoder was used for data representation for understanding semantic relationships between images and text. Like autoencoder, CLIP encoder involve encoding data into a compressed representation. Their framework's experimental results are greatly impacted by the CLIP encoder representation, which allows for deeper interaction detection, better multimodal fusion, better feature learning, and fine-tuning advantages. In order to improve the system's capacity to extract significant features, capture semantic relationships, and enhance overall recommendation system performance in the context of news recommendation.

2.2. Drug discovery and deep learning

Deep learning models can analyze large datasets of chemical and biological information to identify patterns [16] and predict which compounds are most likely to have therapeutic effects. These models can be trained on a variety of data sources, including chemical structures, genomic data, and clinical trial results.

By integrating a deep generative model, kinase selectivity screening, and molecular docking, deep learning methodologies were reported in [17] as a scaffold-based molecular design methodology, leading to a new discoidin domain receptor1 (DDR1) inhibitor compound 2 that demonstrated a robust DDR1 inhibitory profile.

A deep recurrent neural network (RNN) and a multitask deep neural network (MTDNN) were used in the automated approach of Xiaoqin et al.'s [18] to generate and optimize multitarget antipsychotic drugs. Their model was successfully capable of identifying high-scoring compounds.

Hamza, H. et al. [19], investigated the use of a deep learning convolutional network that enables the prediction of molecular ligand-based targets and their bioactivities using a novel molecular matrix representation. A novel Molecule to functional groups (Mol2Fgs) technique was used for data representation. They resulted that the CNN-QSAR algorithm showed high prediction rates with accuracy 90.21%.

A deep neural network model was proposed by Carvalho, F.G, et al. in [20] to predict the effect of anticancer drugs in tumors through IC50. They pre-trained autoencoders with high dimensional gene expression and mutation data to capture crucial features of tumors, which are then translated to cancer cell lines in order to predict the impact of genetic variations on a given drug. SMILES structures are also included in their model to capture relevant features regarding the drug compound and uses drug sensitivity data, correlated with genomic and drugs data, to identify features that predict the IC50 value for each pair of drug-cell line. A performance of mean squared error of 1.07 resulted the effectiveness of their extracted deep representations in the prediction of drug-target interactions.

Using machine learning approaches, a COVID-19 drug target prediction model was presented by Zamilalo, A., et al. in [21]. The performance of three prediction models was analyzed to predict drug-target docking scores. 300,457 molecules were predicted on 18 different COVID-19 protein docking targets, resulted a competitive performance with $R^2 = 0.69$, MAE = 0.285 and MSE = 0.627.

She, S., et al. [22] presented a deep learning-based model that predicts novel synergistic multi-drug combinations in a given cell line. The model consisted of fully connected feed forward deep neural network. Their model achieved a high performance in regression with MSE = 2.5 and RMSE = 1.58, in addition the classification accuracy was 94%.

The recommendation system proposed in [23] can be adapted and utilized in drug discovery processes to enhance decision-making and optimize research efforts. This system was a novel hierarchical Bayesian model that integrates social network structure through matrix factorization and item content information via latent dirichlet allocation (LDA) for item recommendation. This model can be employed to predict potential interactions between drugs and biological targets based on similarities in drug properties, target characteristics, and social network information of researchers or experts in the field.

Haotong Qin, et al. [24] proposed another system called IR-QLoRA for accurate LoRA-Finetuning Quantization of Large Language Models (LLMs) that can be used in drug discovery. Large Language Models (LLMs) can be used for text mining and analysis of vast amounts of biomedical literature, patents, and clinical trial data. By accurately quantizing LLMs using techniques like IR-QLoRA, researchers can improve the efficiency and accuracy of extracting relevant information for drug discovery. LLMs was also used by [25].

Table 1 present a summary of drug discovery using deep learning models. This table includes the author, research objective, dataset used, data representation, representation algorithm, representation size, prediction algorithm, results method, and Value.

Most of the work presented in literature usually consider data representation with vector size between 128 and 384 as shown in [19, 20]. However, working on a smaller data representation can capture the most important information while discarding noise. In addition it can be computationally efficient, as they require fewer computations and memory. This might be important aspect while working with resource-constrained environments.

As to the far of our knowledge the maximum achieved accuracy in means of MSE, MAE and precision were 2.5, 0.21 and 94%. Hence, there exists a space for results enhancement.

3. The proposed framework

The proposed framework aims to estimate bioactivity of compounds for a specific protein. The architecture of the proposed framework consisted of four main stages as shown in Fig. 5. The first one is pre-processing stage; given the SMILES data set, tokenization took place then padding step used to set all the tokenized SMILES to the same length, which was 100 in both data sets. This length was the maximum smile length occurred in the dataset. The second stage was the autoencoder which consisted of six layers. Three layers in encoder phase which were input layer, embedding layer and LSTM layer, and three layers in decoder phase which were repeated vector layer, LSTM layer and Time distributed dense layer. The encoder output latent space vector with length sixty-four [26]. Adam optimizer was used as stochastic gradient descent method. Adam optimization involves the use of adaptive learning rates which helps to improve the accuracy of the model. This optimizer adjusts the learning rate based on the parameters of the model, allowing it to reach a global minimum more quickly. This helps to reduce the amount of time it takes to train the model, as well as improve its accuracy [27]. The output latent space will be classified using Feed Forward architecture. The classifier will output one of three classes BAC0 class which means Bioactivity less than 10, or BAC1 class which means Bioactivity between 10 and 100, or BAC2

Table 1
Drug Discovery with Deep learning models.

Paper	Research Objective	Dataset Used	Data Representation	Representation Algorithm	Representation Size	Prediction Algorithm	Results Method	Results Value
Hamza, H., et al [19]	Bioactivity Prediction	Chembl	Compound Structure	Mol2Fgs	128	Fully Connected Layers with one Drop out layer	Accuracy	90.21%
Carvalho, F.G., et al [20]	Predicting IC50	Genomics of Drug Sensitivity in Cancer database	GENE Expressions, Mutation and SMILES	Autoencoder	384	Regularized Feed Forward Predictor	MSE	1.07
Zamitalo, A., et al [21]	Binding Affinity Prediction	ENAMINE, ZINC, and Drug Bank	SMILES	Fingerprint Encoding	166	XGBoost Regression	MAE	0.21
She, S., et al [22]	Predict novel synergistic multi-drug	Genomics of Drug Sensitivity in Cancer Drug Bank	Genomic Information & Drug-Target Information	Integration	215-dimensional genomic & 1093-dimensional target	Classification & Regression	Accuracy & MSE	94% & 2.5

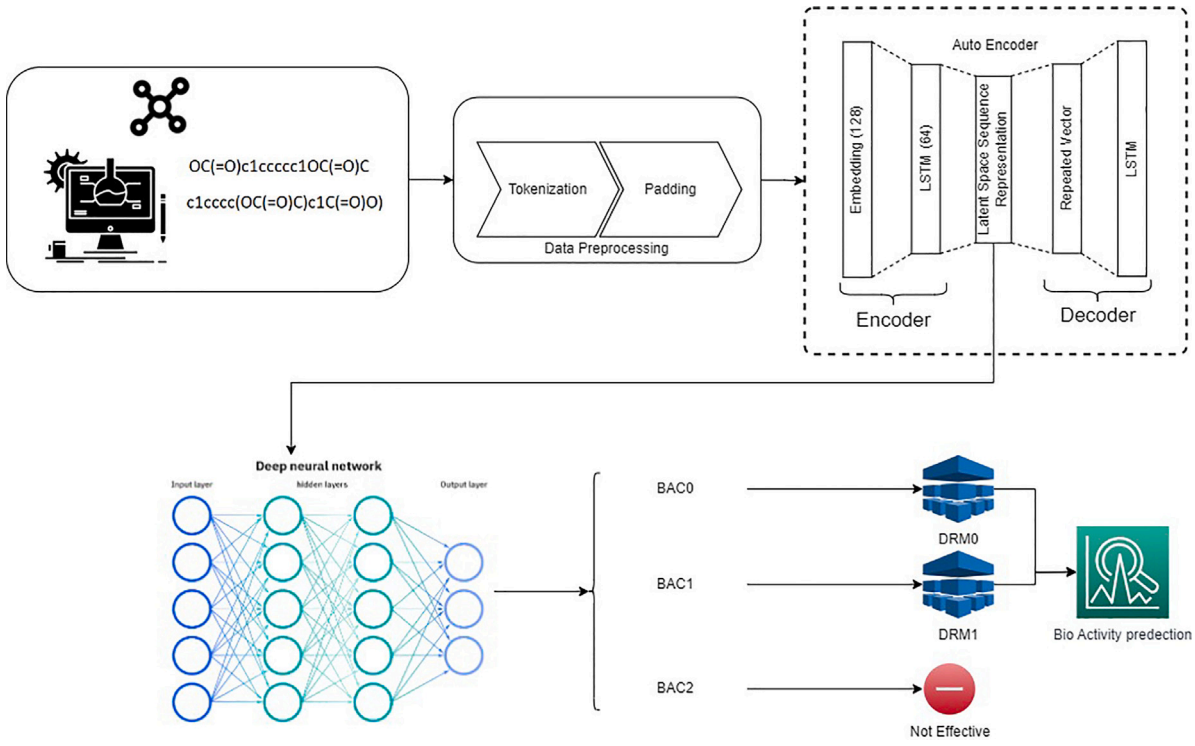


Fig. 5. Proposed system pipeline.

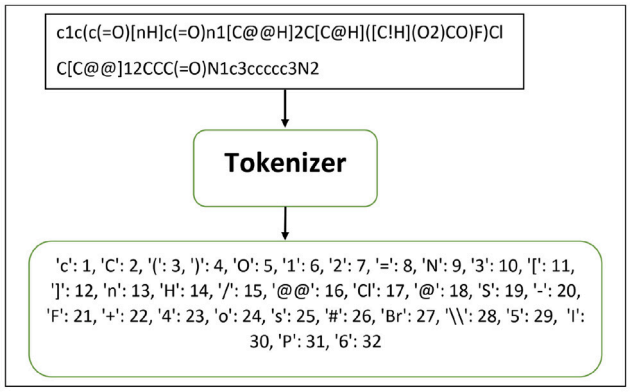
which means Bioactivity greater than 100 (non effective bioactivity). A customized pretrained deep regression model (DRM[0-1]) took place to estimate an accurate value of the bioactivity. Each phase of the proposed framework is detailed in the following subsections.

3.1. Data representation

AutoEncoder (AE) is among the different methods used for data representation. The main aim of the AE is to construct a latent space of compressed representation with low dimension, in which the reconstruction of each element to the original input is attained. The encoder is the part of the framework that considers the original data as input, which has a lot of dimensions, and condenses it into a low-dimensional format (latent space). The decoder then process this latent space representation and reconstruct the original data representation [28,29].

Data Preprocessing comprises of two steps, as shown in Fig. 5. The first step is "Tokenizer" and then "Padding". A tokenizer was used to assign a number to each character in the smile dictionary, so all of the smile samples can be transformed into numbers. Fig. 6(a) shows the representation of the tokenizer in which each number represents an element, while Fig. 6(b) shows SMILES samples following representation of Tokenizers. Next, "Padding" was utilized to add elements at the end of each input with a same value resulting all the inputs in same size.

Habitually, using neural networks like RNNs, LSTMs, and CNNs, the encoder and decoder are set up. Since SMILES data is in sequence form, LSTM was used for AE specifically because it was designed to process data of that kind [5,30]. Before using an Encoder LSTM,



(a)

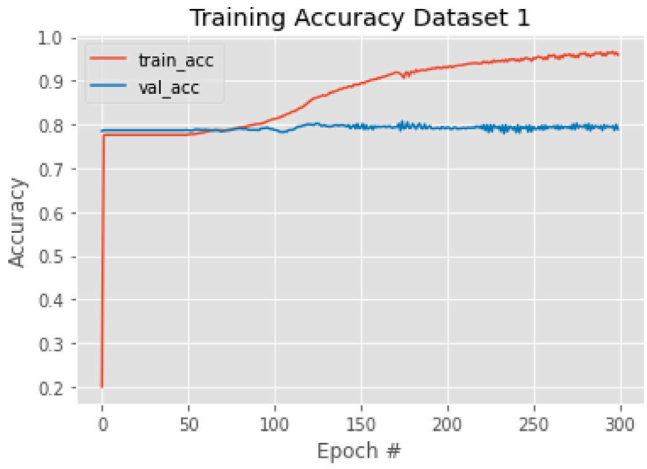
Smile	Smile After tokenizer
C[C@@]12CCC(=O)N1c3ccccc3N2	[2,11, 2,16,12, 6, 7,2, 2, 2, 3, 8, 5,4,9, 6,1, 10, 1, 1, 1, 1, 1, 10, 9, 7]
Cc1c(oc(n1)c2ccc(cc2)Cl)COC(C)(C)C(=O)[O-]	[2, 1, 6, 1, 3, 24, 1, 3, 13, 6, 4, 1, 7, 1, 1, 1, 3, 1, 1, 7, 4, 17, 4, 2, 5, 2, 3, 2, 4, 3, 2, 4, 2, 3, 8, 5, 4, 11, 5, 20, 12]
c1cc(c(cc1F)F)C(Cn2cncn2)(Cn3cncn3)O	[1, 6, 1, 1, 3, 1, 3, 1, 1, 6, 21, 4, 21, 4, 2, 3, 2, 13, 7, 1, 13, 1, 13, 7, 4, 3, 2, 13, 10, 1, 13, 1, 13, 10, 4, 5]

(b)

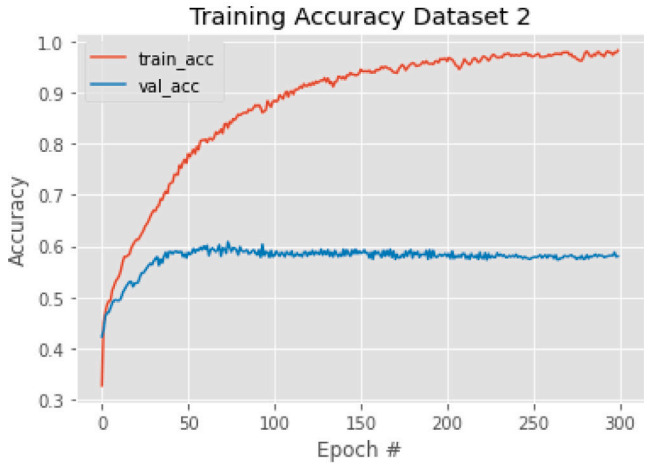
Fig. 6. (a) Tokenizer representation, (b) Smiles after Tokenization.

Smile After Tokenizer	Smile after Embedding
[2,11, 2,16,12, 6, 7,2, 2, 2, 3, 8, 5,4,9, 6,1, 10, 1, 1, 1, 1, 1, 10, 9, 7]	[[0.8,0.3,0.4,0.9,…..], [0.1,0.2,0.4,0.3,…..], [0.8,0.3,0.4,0.9,…..], [0.9,0.2,0.4,0.3,…..], ……..]

Fig. 7. Embedding layer output.



(a)



(b)

Fig. 8. (a) Training Accuracy Dataset 1, (b) Training Accuracy Dataset 2.

positive integers are converted to dense vectors having a fixed size via an Embedding layer. Fig. 7 shows the tokenized SMILE after using Embedding layer. Using a repeated vector layer as an adapter to fit the encoder's fixed-sized 2D output to the decoder's desired 3D input, which has a different length. Additionally, a Time distributed layer was used to enable the reuse of the same output layer for each component of the output sequence. The next step will involve extracting the latent space as a feature vector.

3.2. Classification model

Different classes of common properties, and what distinguishes them, are expected to be discovered by the framework, leading to predict new unseen cases classification correctly [21]. Feed Forward Neural Network is used as classification algorithm in order to predict bioactivity class.

The used Neural Network was composed of two hidden layers and an output layer. The input layer received feature vector with length 64. The proposed framework learnt 128 weights in this layer and applied the relu activation function. Increasing the dimensionality of the feed forward neural network was necessary as far as the model is not capturing the complexity of the data. Moreover, increasing the number of neurons in the hidden layers can help improve the model's performance. This can allow the network to learn more complex relationships between the input features and the output [31]. The next layer learnt 64 weights. Finally, a fully connected layer of 3 weights corresponding to (BAC0, BAC1, BAC2) output classes. The three classes of bioactivity were established; "BAC0" represents bioactivity values less than 10, the second, those between 10 and 100 named as "BAC1", and the third, those greater than 100 named as "BAC2". The softmax function was used as threshold in the study. Softmax activation, which works well [32] with multiclass classification, is utilized in the output layer of the defined model, which is a multiclass network model with ten output classes. The softmax function was used in similar drug discovery [33]. Empirically, SGD with three constant learning rates $\eta = 0.1$, $\eta = 0.005$, and $\eta = 0.001$ was experimented, and observed that $\eta = 0.1$ is too large; $\eta = 0.001$ is too small; and $\eta = 0.005$ tends to be good for the tested networks. A learning rate that is too high may cause the optimizer to overshoot the minimum, leading to oscillations or divergence, while a learning rate that is too low may result in slow convergence or getting stuck in a suboptimal solution [34]. The neural network used the category cross-entropy loss metric while showing the accuracy. The network was trained for a total of 300 epochs. Fig. 5 shows the architecture for the implemented NN. Fig. 8(a) shows the Classification training and validation accuracy results experimenting Cyclooxygenase-2 target protein dataset while Fig. 8(b) shows the training and validation accuracy results of Carbonic anhydrase XII target protein dataset.

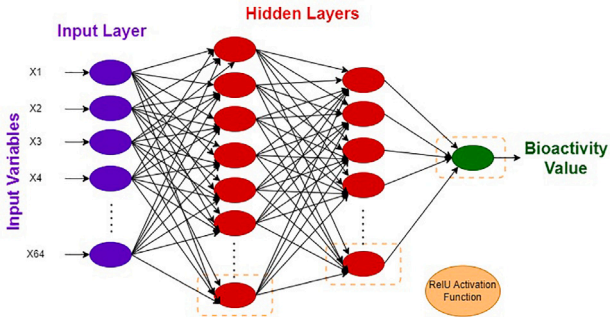


Fig. 9. Deep regression model.

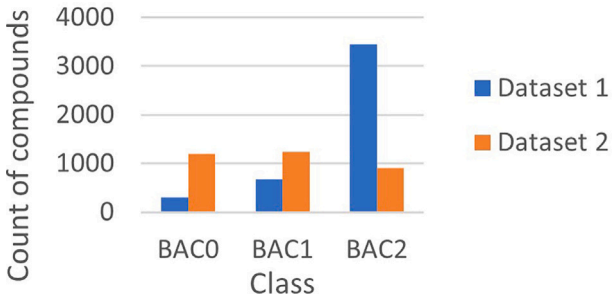


Fig. 10. Distribution of datasets.

3.3. Deep regression model

Prediction of continuous values is the goal of tasks to be solved through regression techniques. An enormous bundle of applicative scenarios is spanned through regression techniques. For example, age estimation [7,8], facial landmark detection [35,36], head-pose estimation [37,38] image registration [39,40], or human pose estimation [41,42]. Deep learning architectures have exceeded the advanced traditional computer vision tasks during the last decade, such as object detection [43,44] or image classification [45,46]. These respective architectures comprise various convolutional layers that are proceeded by layers that fully connected, and softmax layer classified with a loss of cross-entropy, for example. Convolutional neural network (CopyNets) is referred to the overall architecture. Likewise classification, regression problems are solved by CopyNets. Within the same paradigm, a fully connected regression layer replaces the SoftMax layer with sigmoid, relu or linear activation functions. This kind of architecture is known as vanilla deep regression.

A fully connected regression model was designed in order to predict the compound bioactivity value. As shown in Fig. 9, this model consisted of two hidden layers and an output layer. The input layer received feature vector with length 64. The proposed regression model learnt 128 weights in this layer and applied the RelU activation function. Finally, a fully connected layer of one output of predicted bioactivity value.

4. Performance evaluation

The objective of the proposed framework is to reliably estimate chemical bioactivity. Evaluation measures the performance, identifies the potential problems, optimizes the hyperparameters, and fine-tunes for achieving best results of bioactivity estimation. Data is separated into training, validation, and test sets, and metrics such as accuracy and reconstruction error score are used to assess performance. This contributes to the development of high-quality predictive models for drug discovery research.

For the purpose of this study three experiments have been conducted to evaluate the performance of the proposed framework on two different datasets.

4.1. Datasets

The proposed framework was experimented on Cyclo-oxygenase-2 target protein dataset (Dataset 1) and Carbonic anhydrase XII target protein dataset (Dataset 2). Both datasets were extracted from CHEMBL database. Dataset 1 consisted of 4433 compounds. While dataset 2 consisted of 3354 compounds. Each compound has its own bioactivity value. For classification, the data was divided into three classes according to their bioactivity value. The first class represents bioactivity values less than 10 named as “BAC0”, the second, those between 10 and 100 named as “BAC1”, and the third, those greater than 100 named as “BAC2”. Fig. 10 shows the histogram of class distribution of dataset 1 and dataset 2.

4.2. Experiments details

This subsection is divided into three parts. Part one shows the results of data representation using autoencoder, then part two discuss the results when applying classification, then part three shows regression model results.

4.2.1. Experiment 1: Autoencoder for data representation

The objective of this experiment is to measure the accuracy of the autoencoder representing the SMILE data. In this experiment we have trained two datasets using autoencoder. Three different evaluation metrics for similarity after reconstructing the output text were calculated. The three similarity measurements are accuracy, BLEU, and METEOR. Natural Language Processing (NLP) uses the metric BLEU (Bilingual Evaluation Understudy) to assess the calibre of material that has been automatically translated. The machine-generated text is compared to one or more reference translations, and a score is determined by how much of the machine-generated text matches the reference text. A higher BLEU score, which runs from 0 to 1, indicates better translation quality. In the NLP field, the BLEU score is frequently used as a benchmark for assessing the effectiveness of machine translation systems [47].

Another measure used in Natural Language Processing (NLP) to assess the effectiveness of machine-translated text is called METEOR (Measure for Evaluation of Translation with Explicit Ordering). Compared to BLUE and ROUGE, it is a more complicated metric because it considers word order, synonyms, and paraphrasing in addition to the degree to which the machine-generated text and the reference text overlap. A higher score indicates higher-quality machine-generated text. The score goes from 0 to 1. In comparison to BLEU, METEOR is thought to be more accurate and robust because it can manage a wider range of linguistic phenomena and is less impacted by problems like word order and sentence structure [48].

The equation used for calculating the BLEU score is presented in Eq. (1) where Geometric Average Precision Score (GAPS) is presented in Eq. (2) where w_n is uniform weight and p_n is the precision score; Brevity Penalty is presented in Eq. (3) where c is the predicted length and r is the target length. The accuracy is measured by evaluating the reconstruction quality using Eq. (4).

$$BLEU(N) = BrevityPenalty.GAPS(N) \tag{1}$$

$$GAPS(N) = exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{2}$$

$$BrevityPenalty = \begin{cases} 1 & ,if \quad c > r \\ \left(e^{1-\frac{r}{c}}\right) & ,if \quad c \le r \end{cases} \tag{3}$$

$$Accuracy = \left(\frac{TP+TN}{TP+TN+FP+FN}\right) * 100 \tag{4}$$

The average BLEU score, average METEOR score and accuracy percentage between SMILE strings and the autoencoder predicted SMILE

Table 2
Autoencoder Results.

	Average BLEU score	Average METEOR score	Accuracy
Dataset 1	0.71	0.66	89%
Dataset 2	0.75	0.71	90%

strings were calculated for Cyclooxygenase-2 target protein dataset and Carbonic anhydrase XII target protein dataset. Table 2 shows the results for BLEU, METEOR and similarity. The BLEU, and METEOR results were between 0 and 1 where indicated a better quality of machine-generated text.

4.2.2. Experiment 2: Classification of three classes

The main objective of this experiment is to classify the data whether it is below 10 or between 10 and 100 or above 100 using Feed forward Neural Network. In this experiment we tested the two datasets with a cross folding of 8 k-folds. Softmax activation function with categorical cross entropy was used.

For evaluating the performance of our model, accuracy, precision, recall and F1-score were calculated. The accuracy has been calculated as shown in Eq. (4). Precision is presented in Eq. (5) where it measures the accuracy of positive predictions, reflecting the ratio of true positives to all positive predictions made by the model. While as shown in Eq. (6), recall evaluates the model's ability to predict all positive instances, indicating the ratio of true positives to all actual positives in the dataset. F1-score as presented in Eq. (7) calculates the harmonic mean of precision and recall, provides a balanced measure of a model's performance, particularly useful when there is an imbalance between the classes in the dataset [49].

The average accuracy of training dataset 1 was 95.73% with precision 91.7%, recall 89.4% and 90.5% F1-score while for dataset 2 was 96.91% with precision 96.3%, recall 95.9% and 96.1% F1-score. The average accuracy of testing dataset 1 was 79.01% with precision 75.1%, recall 73.3% and 74.18% F1-score while for dataset 2 was 60.26% with precision 74.2%, recall 73.6% and 73.9% F1-score.

Moreover, we have applied a single way analysis of variance (ANOVA). The results of training showed that there were significant differences across the 8 k-folds. The p-value was less than 0.05 (3.79E−06), F-crit (4.6) was less than overall F value (53.57). Also, the results of testing showed that there were significant differences across the 8 k-folds. The p-value was less than 0.05 (2.16E−11), F-crit (4.6) was less than overall F value (360.78).

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{7}$$

4.2.3. Experiment 3: Deep regression model

The main objective of this model was to predict bioactivity value using Deep regression model DRM0 and DRM1 where DRM0 was experimented on BAC0 data (bioactivity less than 10) in each dataset; and DRM1 was experimented on BAC1 data (bioactivity between 10 and 100) in each dataset.

Model was experimented in each dataset twice, one for the data with bioactivity less than 10 and the other for the ones whose bioactivity between 10 and 100. The evaluation metrics used for evaluating the predicted values were "Mean Absolute Error" (MAE) as shown in Eq. (8), "Mean Square Error" (MSE) as presented in Eq. (9) and "Route Mean Square Error" (RMSE) as presented in Eq. (10). Where

Table 3
Regression Results in BAC0.

	BAC0					
	Train			Test		
	MAE	MSE	RMSE	MAE	MSE	RMSE
Dataset1	0.7	0.8	0.89	2.96	12.9	3.56
Dataset2	0.46	0.56	0.79	2.44	10.82	3.29

Table 4
Regression Results in BAC1.

	BAC1					
	Train			Test		
	MAE	MSE	RMSE	MAE	MSE	RMSE
Dataset 1	0.9	2.98	1.73	40.1	2536.2	50.36
Dataset 2	0.19	0.64	0.8	37.9	2403.05	49.03

y_i is the predicted value, x_i is the true value and n is the total number of data points.

$$MAE = \sum_{i=1}^{n} \frac{|y_i - x_i|}{n} \tag{8}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i)^2 \tag{9}$$

$$RMSE = \sqrt{MSE} \tag{10}$$

In this experiment we have tested two datasets. RelU activation function with mean absolute error was used. Tables 3 and 4 show the results of training and testing in each data.

5. Discussion

The ability of autoencoders to learn a continuous latent representation makes it simpler to employ them for downstream tasks like classification and regression. This is because, in contrast to discrete representations, continuous representations can be easily modified using gradient-based optimization.

During initial experiments, deep regression model without classification was found to perform extremely poor. Excluding the non-effective bioactivity where values are above 100 and clustering of the data to three classes BAC0, BAC1 and BAC2 were found to be effective steps to process the data before feeding it to the regression model.

Our concept of classification followed by regression was found to be an effective method for forecasting biological activity. The regression model predicts the potency of the active molecules after the categorization model determines whether a molecule is active or inactive. This method enables us to capture both the continuous nature of potency and the binary aspect of biological activity (active vs. inert).

Most of the studies reviewed in the related work are based on secondary data analysis providing insights into previous studies related to the data representation and drug discovery such as [13,16–18].

In [14], the researchers used a dataset with limited molecular complexity, which may not fully capture the diversity of compounds and bioactivity values. This limitation is overcome in the current study as Chembl dataset contains a wide set collection of bioactivity data.

The work presented by [20] was applied on anticancer protein target only. However, our proposed framework can be applied on any protein target. The main difference in the predicted results was related to different protein targets.

6. Conclusion and future work

This paper has presented a novel framework for data representation and predicting of compound bioactivity through presenting the integration of autoencoders and deep learning methodologies. The latent space generated by autoencoder for SMILE representation was utilized by classification then regression models. A feed forward neural network classified the latent space output one of three classes bioactivity less than 10, or bioactivity between 10 and 100, or bioactivity greater than 100 (non effective bioactivity). According to the classified class, an accurate biological activity value was predicted using a customized pretrained deep regression model. The results of this paper showed that the regression model can predict the bioactivity with high effectiveness.

The scalability of autoencoder architecture was one of our limitations or constraints encountered during the implementation of our proposed model, particularly when dealing with larger datasets or more complex molecular structures. Additionally, deep regression model with-out classification was found to perform extremely poor during initial experiments. In future work, incorporating transformer-based models looks like a great way to enhance chemical bioactivity prediction. Through the adaptation of transformer architectures to represent SMILES, researchers may be able to increase the accuracy of bioactivity predictions and the representation learning process. Additionally, leveraging pre-trained transformer models, fine-tuned on large-scale molecular datasets, could facilitate transfer learning and improve the generalization capabilities of bioactivity prediction models [50].

CRediT authorship contribution statement

Yasmine Eid Mahmoud Yousef: Conceptualization, Methodology, Software, Writing – original draft. **Ayman El-Kilany:** Methodology, Supervision, Validation, Writing – review & editing. **Farid Ali:** Supervision, Validation, Writing – review & editing. **Yassin M. Nissan:** Data curation, Investigation, Supervision, Visualization. **Ehab E. Hassanein:** Supervision, Writing – review & editing.

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.

References

[1] Lu P, Bevan DR, Leber A, Hontecillas R, Tubau-Juni N, Bassaganya-Riera J. Computer-aided drug discovery. Accel Path Cures 2018;7–24.

[2] Hu S, Chen P, Gu P, Wang B. A deep learning-based chemical system for QSAR prediction. IEEE J Biomed Health Inform 2020;24(10):3020–8.

[3] Nothias L-F, Nothias-Esposito M, Da Silva R, Wang M, Protsyuk I, Zhang Z, Sarvepalli A, Leyssen P, Touboul D, Costa J, et al. Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation. J Nat Prod 2018;81(4):758–67.

[4] Barba-Ostria C, Carrera-Pacheco SE, Gonzalez-Pastor R, Heredia-Moya J, Mayorga-Ramos A, Rodríguez-Pólit C, Zúñiga-Miranda J, Arias-Almeida B, Guamán LP. Evaluation of biological activity of natural compounds: Current trends and methods. Molecules 2022;27(14):4490.

[5] Dikshit A, Pradhan B, Alamri AM. Pathways and challenges of the application of artificial intelligence to geohazards modelling. Gondwana Res 2021;100:290–301.

[6] Idakwo G, Thangapandian S, Luttrell IV J, Zhou Z, Zhang C, Gong P. Deep learning-based structure-activity relationship modeling for multi-category toxicity classification: a case study of 10k tox21 chemicals with high-throughput cell-based androgen receptor bioassay data. Front Physiol 2019;10:1044.

[7] Ucak UV, Ashyrmamatov I, Lee J. Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. J Cheminformatics 2023;15(1):55.

[8] Wigh DS, Goodman JM, Lapkin AA. A review of molecular representation in the age of machine learning. Wiley Interdiscip Rev Comput Mol Sci 2022;12(5):e1603.

[9] Li C, Feng J, Liu S, Yao J, et al. A novel molecular representation learning for molecular property prediction with a multiple SMILES-based augmentation. Comput Intell Neurosci 2022;2022.

[10] Tetko IV, Engkvist O. From big data to artificial intelligence: chemoinformatics meets new challenges. J Cheminformatics 2020;12:1–3.

[11] Niazi SK, Mariam Z. Recent advances in machine-learning-based chemoinformatics: A comprehensive review. Int J Mol Sci 2023;24(14). http://dx.doi.org/10.3390/ijms241411488, URL https://www.mdpi.com/1422-0067/24/14/11488.

[12] Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. J Materiomics 2017;3(3):159–77.

[13] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science 2018;4(2):268–76.

[14] Bjerrum EJ, Sattarov B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. Biomolecules 2018;8(4):131.

[15] Han S, Huang W, Luan X. VLSNR:Vision-linguistics coordination time sequence-aware news recommendation. 2022, arXiv:2210.02946.

[16] Nag S, Baidya AT, Mandal A, Mathew AT, Das B, Devi B, Kumar R. Deep learning tools for advancing drug discovery and development. 3 Biotech 2022;12(5):110.

[17] Tan X, Li C, Yang R, Zhao S, Li F, Li X, Chen L, Wan X, Liu X, Yang T, et al. Discovery of pyrazolo [3, 4-d] pyridazinone derivatives as selective DDR1 inhibitors via deep learning based design, synthesis, and biological evaluation. J Med Chem 2021;65(1):103–19.

[18] Tan X, Jiang X, He Y, Zhong F, Li X, Xiong Z, Li Z, Liu X, Cui C, Zhao Q, et al. Automated design and optimization of multitarget schizophrenia drug candidates by deep learning. Eur J Med Chem 2020;204:112572.

[19] Hamza H, Nasser M, Salim N, Saeed F. Bioactivity prediction using convolutional neural network. In: Emerging trends in intelligent computing and informatics: data science, intelligent information systems and smart computing 4. Springer; 2020, p. 341–51.

[20] Carvalho FG, Abbasi M, Ribeiro B, Arrais JP. Deep model for anticancer drug response through genomic profiles and compound structures. In: 2022 IEEE 35th international symposium on computer-based medical systems. CBMS, IEEE; 2022, p. 1–6.

[21] Zamitalo A, Xie Q, Allam M, Philip P, Shi W, Giuste F, Marteau B, Murakoso M, Wang MD. Development of machine learning regression model for COVID-19 drug target prediction. In: 2022 IEEE international conference on bioinformatics and biomedicine. BIBM, IEEE; 2022, p. 2808–15.

[22] She S, Chen H, Ji W, Sun M, Cheng J, Rui M, Feng C. Deep learning-based multi-drug synergy prediction model for individually tailored anti-cancer therapies. Front. Pharmacol. 2022;13:1032875.

[23] Purushotham S, Liu Y, Kuo CCJ. Collaborative topic regression with social matrix factorization for recommendation systems. 2012, arXiv:1206.4684.

[24] Qin H, Ma X, Zheng X, Li X, Zhang Y, Liu S, Luo J, Liu X, Magno M. Accurate lora-finetuning quantization of LLMs via information retention. 2024, arXiv:2402.05445.

[25] Huang W, Liu Y, Qin H, Li Y, Zhang S, Liu X, Magno M, Qi X. Billm: Pushing the limit of post-training quantization for LLMs. 2024, arXiv:2402.04291.

[26] Cavallari GB, Ribeiro LS, Ponti MA. Unsupervised representation learning using convolutional and stacked auto-encoders: a domain and cross-domain feature space analysis. In: 2018 31st SIBGRAPI conference on graphics, patterns and images. SIBGRAPI, IEEE; 2018, p. 440–6.

[27] Zhang Z. Improved adam optimizer for deep neural networks. In: 2018 IEEE/ACM 26th international symposium on quality of service. Ieee; 2018, p. 1–2.

[28] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT Press; 2016.

[29] Bian Y, Xie X-Q. Generative chemistry: drug discovery with deep learning generative models. J. Mol. Model. 2021;27:1–18.

[30] Du S, Li T, Yang Y, Gong X, Horng S-J. An LSTM based encoder-decoder model for MultiStep traffic flow prediction. In: 2019 international joint conference on neural networks. IJCNN, IEEE; 2019, p. 1–8.

[31] Thomas AJ, Petridis M, Walters SD, Gheytassi SM, Morgan RE. On predicting the optimal number of hidden nodes. In: 2015 international conference on computational science and computational intelligence. CSCI, IEEE; 2015, p. 565–70.

[32] Waoo AA, Soni BK. Performance analysis of sigmoid and relu activation functions in deep neural network. In: Intelligent systems: proceedings of SCIS 2021. Springer; 2021, p. 39–52.

[33] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today 2018;23(6):1241–50.

[34] Senior A, Heigold G, Ranzato M, Yang K. An empirical study of learning rates in deep neural networks for speech recognition. In: 2013 IEEE international conference on acoustics, speech and signal processing. IEEE; 2013, p. 6724–8.

[35] Wu Y, Ji Q. Facial landmark detection: A literature survey. Int J Comput Vis 2019;127:115–42.

[36] Lv J, Shao X, Xing J, Cheng C, Zhou X. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 3317–26.

[37] Abate AF, Barra P, Pero C, Tucci M. Head pose estimation by regression algorithm. Pattern Recognit Lett 2020;140:179–85.
[38] Huang B, Chen R, Xu W, Zhou Q. Improving head pose estimation using two-stage ensembles with top-k regression. Image Vis Comput 2020;93:103827.
[39] Sokooti H, Saygili G, Glocker B, Lelieveldt BP, Staring M. Quantitative error prediction of medical image registration using regression forests. Med Image Anal 2019;56:110–21.
[40] Cao X, Yang J, Zhang J, Wang Q, Yap P-T, Shen D. Deformable image registration using a cue-aware deep regression network. IEEE Trans Biomed Eng 2018;65(9):1900–11.
[41] Zhao L, Peng X, Tian Y, Kapadia M, Metaxas DN. Semantic graph convolutional networks for 3d human pose regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 3425–35.
[42] Moreno-Noguer F. 3D human pose estimation from a single image via distance matrix regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 2823–32.
[43] Harakeh A, Waslander SL. Estimating and evaluating regression predictive uncertainty in deep object detectors. 2021, arXiv preprint arXiv:2101.05036.
[44] Zou Z, Chen K, Shi Z, Guo Y, Ye J. Object detection in 20 years: A survey. Proc IEEE 2023.
[45] Wang W, Liang D, Chen Q, Iwamoto Y, Han X-H, Zhang Q, Hu H, Lin L, Chen Y-W. Medical image classification using deep learning. Deep Learn Healthc Paradigms Appl 2020;33–51.
[46] Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. 2017, arXiv preprint arXiv:1712.04621.
[47] Reiter E. A structured review of the validity of BLEU. Comput Linguist 2018;44(3):393–401.
[48] Saadany H, Orasan C. BLEU, METEOR, bertscore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. 2021, arXiv preprint arXiv:2109.14250.
[49] Vujovic Z. Classification model evaluation metrics. Int J Adv Comput Sci Appl 2021;Volume 12:599–606. http://dx.doi.org/10.14569/IJACSA.2021.0120670.
[50] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. 2023, arXiv:1706.03762.