Revue d'Intelligence Artificielle
Vol. 38, No. 5, October, 2024, pp. 1381-1389
Journal homepage: http://iieta.org/journals/ria

Prediction of Mechanism of Action Using Ensembled Deep Neural Networks Splits

Rehan Ullah Khan

Department of Information Technology, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia

Corresponding Author Email: re.khan@qu.edu.sa

https://doi.org/10.18280/ria.380502

Received: 27 April 2024
Revised: 25 June 2024
Accepted: 10 July 2024
Available online: 23 October 2024

Keywords:
deep learning, Machine Learning, drug discovery, artificial neural network

ABSTRACT

Mechanisms of Action (MoA) relates to how drugs work at a molecular level to produce therapeutic effects in the body. Every drug interacts with specific molecules and these interactions can either boost or hinder the activities of these molecules, leading to changes in how cells and tissue's function. This article proposes a network of several Ensembled Deep Neural Networks Splits (EDNNS) to predict the multiple targets of the MoA responses of different samples. The dataset consists of various groups of features, with more than two hundred enzyme and receptor targets. Several Machine Learning (ML) models, including the EDNNS, are evaluated. For performance evaluation, the logarithmic loss function is used. In the experimental evaluation, we evaluated MLP, Deep NN, ResNet, Xgboost and found that the proposed ensemble EDNNS is more robust than the MLP, Deep NN, ResNet, Xgboost having less loss. The comparatively less loss thus means a more robust and accurate model for prediction. This work can benefit the advanced drug discovery cause-effect by providing valuable insights and exciting directions for future research.

1. INTRODUCTION

Mechanisms of Action (MoA) is a phenomenon of how drugs work at a molecular level to produce their therapeutic effects in the body. As drugs interact with specific molecules, like receptors or enzymes, they play important roles in biological responses and processes. These interactions augment the activities of these molecules, leading to changes in how cells and tissue's function. Knowing a drug's MoA is vital in pharmacology because it allows researchers to predict how effective it will be, what side effects might occur, and how it might interact with other medications. Machine Learning (ML) provides a valuable contribution to several fields ranging from engineering, social, computing, and medical sciences. Interest has grown in the applicability of the ML in medicine and biosciences, signifying our interest in this field. If the ML models can successfully learn the MoA distribution with minimum loss and high accuracy, the obtained model can then predict a compound's MoA based on a specific cellular signature, benefiting the advanced drug discovery cause-effect process. Understanding the process of MOA is crucial in the identification of drug efficacy concentration in addition to the toxic and lethal dose along with any adverse effects.

The current paper proposes a network of EDNNS to predict the multiple targets of the MoA responses of different samples. As such, an exploratory answer to the question is reported! Can the drug's MoA based on gene expression and cell viability data be predicted using the ML paradigm? The Harvard's Laboratory for Innovation Science dataset is used to explore the answer. The dataset consists of various groups of features with more than two hundred targets of enzymes and receptors. The samples are profiled at different time points and doses. Multiple ML models are evaluated with special interest inclined towards the EDNNS model due to its superior regression performance. Our proposed approach is similar to the stacking ensemble of multiple deep NNs. For performance evaluation, the average value of the logarithmic loss function is used. The experimental evaluation shows that the ML paradigm, especially the EDNNS, provides better MOA prediction and can be used for practical applications and scenarios.

Understanding a drug's MOA is not a pre-requisite for drug approval as long as the safety and efficacy are well documented. Even though largely unclear, Metformin's MOA is proposed to control diabetes through AMP-activated protein kinase's (AMPK) inhibition [1]. Similarly, Dimebon, which was postulated to stabilize mitochondria, thereby possessing anti-Alzheimer potency, nevertheless had to be aborted during the phase 3 trial as cognition potency was due to its histamine and serotonin receptors interaction and not as an anti-Alzheimer effect [2]. A compound's MOA is attributed to the protein signaling in a pathway, the role of effector protein, and differing protein expressions depending on the site of action [3].

The latest technological research methodologies can comprehensively test any compound's hypothesized MOA at many levels using ML, pathway enrichment, connectivity mapping, causal reasoning, and the compound's previously documented site of action and signaling proteins interaction. The current study provides an overview of the different data levels, and compounds' MoA can be elucidated.

Bioactivity data like High Throughput Screening (HTS) provides resourceful data in predicting sites of MoA for orphan drugs [4] wherein a multitude of drug molecules can be evaluated against a panel of compound targets. The obtained data has its limitations as it does not consider the pharmacokinetic properties of absorption, distribution, metabolism, and excretion (ADME) and lacks data at the cellular level about the modified signaling pathway in response to drug binding on the receptors. Many other freely accessible databases like ChEMBL, PubChem, ExCAPE, and BindingDB extract the data from the published literature with different lab and experimental settings and assays. The variation in experiments may be related to a data error, adding to the complexity of the generated data. The morphological modifications like apoptosis or alteration in cytoskeletal protein upon adding any compound to the cell line can be visualized using cell imaging [5]. New assays like the Cell Painting assay [6] using fluorescent dyes depict morphological changes in organelles.

There is various segmentation software capable of analyzing and mining features from 2D cell images like CellProfiler [7], CellCognition [8], and PhenoRipper [9], in addition to automated feature mining software like Convolutional Neural Networks (CNNs) [10] that involves deep learning (DL). DL utilizes raw images for extraction and precise identification of cells or cellular sub-compartments or substructures [11, 12]. One limiting factor for DL software is the inability of all tested drug molecules to alter cellular morphology.

Proteomics provides vital information related to the compound's MOA through the expression of the proteins. The methodology's limitations are time-consuming, expensive, and lacking quantification of all proteins in an experiment, which can be circumvented by obtaining data from multiple experiments or studies [13, 14]. Similarly, metabolomics provides changes in metabolite enzyme activity by any drug molecule. Deeper mechanistic information of any drug molecule can thus be obtained through changes in transcription, translation, proteomics, and metabolites levels. The metabolic method's limitations require multiple experiments to obtain the complete metabolome coupled with data interpretation difficulty and wide data variability requiring experimental replication [15]. Phosphoproteomics provides additional potential pathways modulated by the drug molecules following alteration in protein phosphorylation. The limitation in phosphoproteomics is overcome through PhosphoSitePlus [16], wherein the phosphorylation sites are mapped to proteins, thereby providing biological context through disease and pathway annotations.

ML has significantly influenced drug discovery, beginning with quantitative structure-activity relationship (QSAR) modeling. This method used statistical techniques to predict biological activity based on chemical structure, despite computational and data limitations at the time. The introduction of high-throughput screening (HTS) and genomics resulted in a surge of biological data, which increased the application of ML algorithms like support vector machines (SVMs) and random forests. These algorithms enabled the rapid analysis of large datasets, facilitating the identification of potential drug targets and accelerating early drug discovery stages. Further advancements in deep learning have greatly advanced the field. For example, AtomNet has accurately predicted small molecule binding affinities to proteins, aiding in the discovery of treatments for diseases such as Ebola and multiple sclerosis. DeepMind's AlphaFold has also revolutionized protein structure prediction. Researchers have used AI to analyze biomedical data, identifying drug repurposing opportunities like baricitinib for COVID-19, which was validated through clinical trials. However, the application of ML in drug discovery faces challenges, including a scarcity of high-quality, labeled data, limited model interpretability, and difficulties in generalizing from training data to real-world scenarios. Addressing these challenges requires strategies such as enhancing data integration across domains, developing interpretable ML models, and using techniques like transfer learning and few-shot learning to tackle data scarcity. Collaborative efforts among multidisciplinary teams are also essential to ensure ML models are grounded in biological insights and practical needs. These efforts promise faster, more efficient, and effective approaches to identifying novel therapeutics and improving patient outcomes.

The article introduces an NN-based Ensemble Deep Neural Network System (EDNNS) for predicting multiple targets of Mode of Action (MoA) responses across various samples. It investigates the prediction of MoA using gene expression and cell viability data within a machine-learning framework. Comparative evaluation with MLP, Deep NN, ResNet, and Xgboost reveals that the proposed EDNNS ensemble exhibits greater robustness with lower loss compared to these individual models. The findings of this article may provide advancements in drug discovery by identifying new therapeutic targets and biomarkers, facilitating personalized medicine, and enabling safer, more effective treatments. It helps create drugs that target specific disease mechanisms, reducing side effects and increasing efficacy. Additionally, we can repurpose existing drugs for new uses, speeding up the development of therapies for conditions lacking effective treatments.

2. BIOLOGICAL NETWORK AND PATHWAY DATA

Proteins, the cellular signaling mediators, are critical in understanding MoA due to their interaction with other proteins, genes, and metabolites. Yeast two-hybrid (Y2H) screening [17] demonstrates interactions absent in in-vivo or affinity purification-mass spectrometry (AP/MS) [18] with an increased percentage of false-positive and negatives are the commonly used experimental methods for protein-protein interaction [19-21]. Various network databases are available through in-house experiments, literature mining, and individual network database compilation [22, 23]. Each network interaction type has its database like STRING [24] (protein-protein), RECON [25] (metabolic), and DoRothEA [26] (TF-gene), and composite networks combining multiple networks like OmniPath [27] and BioGRID [28]. Network selection will depend on the research question to be resolved along with network analysis and types of interactions required.

In MoA evaluation, pathway data links genes/proteins to observed phenotypes, allowing straightforward data interpretation. A compound's MoA can be inferred if any compound for known genes in a given pathway demonstrates differential expression. Pathway database, with its limitation of the varied data curation method, can be resolved using PathMe [29] to determine variations and obtain a consensus pathway or select an all-inclusive annotation database.

3. METHODS OF MOA ELUCIDATION

3.1 Connectivity mapping

Connectivity mapping equates modification of gene expression cell lines incubated with the drug molecule to the set of already present reference signatures associated with any disease or other known drug molecules MOA or disease.

3.2 Pathway enrichment

Pathway enrichment analysis provides the opportunity to reduce a vast array of genes or proteins, even in the absence of any biological information, to more minor processes, making them more interpretable than the gene database itself, thereby providing more context in identifying any given phenotype interest.

3.3 Causal reasoning

Any modification in mRNA expression is determined precisely through causal reasoning using prior knowledge network (PKN) molecular interaction from gene expression data [30]. Any compound's MoA is determined using such methodology through the compound-induced modulated signaling proteins. A compound's MoA is determined accurately without mixing gene level with protein activity through causal reasoning's estimated perturbed protein signaling equated to pathway enrichment. One limitation of casual reasoning is an absence of validated signaling protein output equated to experimentally determined protein activity modification.

3.4 Clustering aggregation

These algorithms encompasses essential tools for exploration, serving the vital purpose of identifying and organizing groups of related or interacting samples [31]. These methods operate based on the principles of similarity or distance, utilizing metrics like k-means clustering or data density measures such as DBSCAN. They play a pivotal role in analyzing complex datasets containing gene expression, chemical compositions, and image-based information, which are often unstructured and high-dimensional [32]. By uncovering patterns and relationships inherent in these intricate datasets, clustering techniques help to unravel the complexity of biological processes. Ultimately, they provide valuable insights that significantly contribute to the advancement of scientific understanding in various fields.

3.5 Group Factor Analysis

Group Factor Analysis (GFA) is an integrative analysis determining the correlation of various data like chemical descriptors and genetic processes [33]. Multi-omics Factor Analysis (MOFA) and MOFA+ are examples of GFA linked to clinically relevant findings in which the integrated clusters comprising 24 anti-tumor compounds' biological data demonstrated to possess MEK inhibitor activity on hematopoietic cells [34].

3.6 Modeling and prediction

Supervised ML trains and identifies patterns in the presence of identified labels used to optimize a function by connecting to features like gene expression with an endpoint like a compound's activity at specific points on the label. Supervised ML demonstrates the compound's MoA by predicting primary and off-target drug interactions [35]. DL methods are typical artificial neural networks (ANN) with a multitude of discrete layers along with specialized training modules, thereby mimicking the human brain's complex neural network.

Each methodology has its benefit and limitations like network and pathway depend on preliminary information related to the curation quality. In addition to being time-constrained, ML has its share of other limitations in interpreting data, and the data obtained is restricted to a small part of the high-level MoA space. Similarly, different data types obtain a varied version of the MoA biology and thus enable a more comprehensive understanding of compound MoA. ML contributes to MOA elucidation through prediction, which upon further interpretation, can be utilized by researchers.

4. PROPOSED APPROACH: ENSEMBLED SPLITS

We target the scenario! Can the drug's MoA based on gene expression and cell viability data using the ensemble stacking as splits be predicted with minimum loss? As such, a model was trained to classify drugs based on the stimulus of the biological activity after the drug's administration. For performance evaluation, the average value of the logarithmic loss function was applied to each drug MoA annotation pair in the dataset.

ML is generally based on the assumption that all algorithms are centralized, implying that both the training data and the model reside on the same computing infrastructure. However, this has recently been considered a barrier in ML innovation concerning data privacy. Therefore, the concept of NN models spread across multiple physical machines has provided an exponential reduction to the computational burden of training and, at the same time, has high accurate models when trained over a large number of spreads. In ensemble stacking splits NN, the training of NN is divided into several sub-trainings performed and distributed to multiple nodes. Each sub-model is a fully functional NN that feeds into the segment. The output of the splits is then concatenated for final prediction and model generation. The details of the splits and the NN are presented in the experimental section.

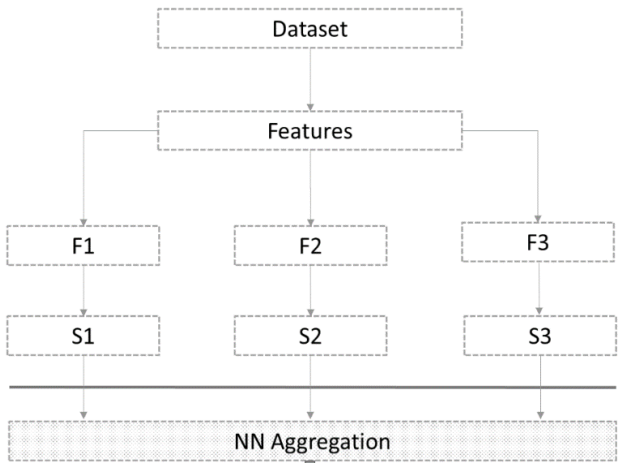

Figure 1. Ensembles as splits for the proposed model

Dataset represents the generic raw form of the data, for example, images. The features represent the extracted features in case data is in raw form. F1, F2, and F3 represent the subset of features where the prominent features are sub-divided. The S1, S2, and S3 represent the independent units of the NN network used for training. Finally, the results are aggregated.

Figure 1 shows the generic flow of the proposed model for the ensemble as splits for the proposed model, wherein the dataset represents the generic raw form of the data, which can be expressed in terms of datasets having images. Figure 1 represents the features extracted from the dataset if the data is in raw form, while if the dataset is already available as features, then the step might be replaced with dataset pre-processing. For ensemble splits, the features are sub-divided as F1, F2, and F3; F represents the subset of features where the prominent features are sub-divided. The S1, S2, and S3 represent the independent units of the NN network used for training. These are physical systems interconnected to solve a particular problem in pure split networks. Finally, the results are aggregated from different nodes.

5. EXPERIMENTAL ANALYSIS

The dataset for MoA prediction is obtained from and available on Kaggle and is provided as such by the Connectivity Map (CM), a project by MIT, Harvard, LISH, and the NIH-LINCS [36]. The dataset is utilized for academic purposes and experimental evaluation of the ML models only. The dataset consists of the following features:

(1) Gene expression
(2) Cell viability
(3) Multiple targets of MoA

The problem represented by MoA data is a multilabel classification problem. Thus, data has multiple targets; however, these are not multiple classes. First, the data analysis is presented, followed by MoA prediction using ensemble stacking.

Table 1 presents a sample of the MoA data. The training features set represents the gene expression data and cell viability and is represented by "g" and "c" symbols. In the dataset, the samples *cp_type* is treated with a *cp_vehicle* compound or with *ctrl_vehicle* (control perturbation). The *ctrl_vehicle* has no MoA. *cp_time* and *cp_dose* represent 24, 48, and 72 hours' duration, and the dose is either high or low.

From a data wrangling point of view, no missing values were found. For visual analysis of the data, Figure 2 shows the plot for the gene expression features, only representing g-0 and g-2. Figure 3 shows the cell viability features, C-0 and C-5, as representative samples.

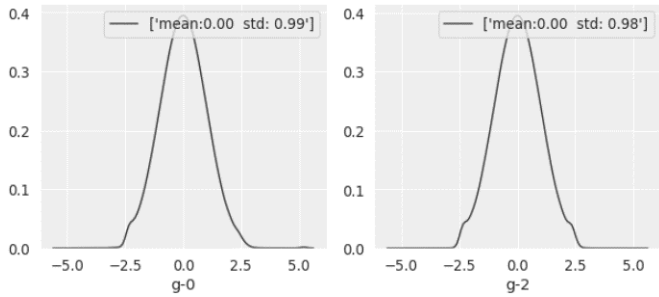

Figure 2. Plot of gene expression features g-0 and g-2

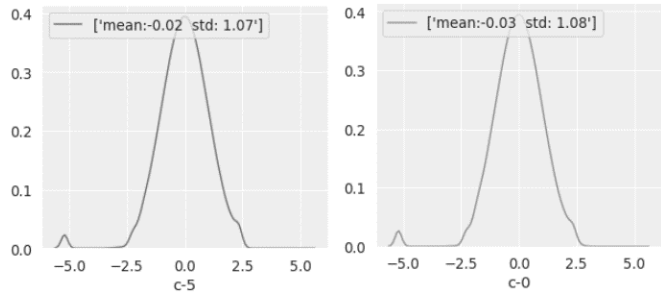

Figure 3. Plot of cell viability features c-0 and c-5

Table 1. Sample of data from the training set

sig_id	cp_type	cp_time	cp_dose	g-0	g-1	g-2	g-3	g-4	g-5
id_000644bb2	trt_cp	24	D1	1.062	0.5577	-0.2479	-0.6208	-0.1944	-1.012
id_000779bfc	trt_cp	72	D1	0.0743	0.4087	0.2291	0.0604	1.019	0.5207
id_000a6266a	trt_cp	48	D1	0.628	0.5817	1.554	-0.0764	-0.0323	1.239
id_107fc335d	ctl_vehicle	48	D2	0.0309	0.4909	1.198	5.031	-1.241	0.4047

6. EVALUATION AND DISCUSSION

Predictions using the proposed EDNNS are performed using the bottom segment, followed by forwarding the prediction to the next model segment. After receiving the prediction, a new prediction is made based on the previous prediction on the input data, which is then forwarded to the next model. These steps are repeated until reaching of end layer with the final prediction and a computational graph for each model. The computational graphs embody input transformation to the prediction and help in the backpropagation phase. Then the network is defined, in this case, a four-layered network where each split is a self-sufficient network. The shape of layers influences the final prediction. The sending and receiving layers must have a similar structure. Next, workers are defined to distribute the splits. Training functions are defined similarly to that of the conventional NN. A second backpropagation phase is needed, which pushes gradients back over the segments. Finally, for starting training, the data is sent to starting locations.

The proposed EDNNS consists of many dense layers, dropout layers, and layers performing batch normalization. After the dense layers, activation functions are altered. For training, the network is trained on the non-scored targets. The weights of the non-scored targets are then transferred for the training of another model on the scored targets. Smoothing on the label is then used, which helps in regularization. For training, the Multilabel Stratified K-Fold with ten splits is adopted. Figures 4 and 5 show the network's architecture. Figure 4 shows the single split and its layers. 85% of the top features are used as input for each split. The Autograd function enables the differentiation of variables versus the loss function. An Autograd gradient can be used to update the model. However, in the implementation, the computational graph

variables were lacking in one place; therefore, the partial backpropagation splits backward loss reduction approach was adopted as splits have to be concatenated to get the final model prediction. The concatenation blending is achieved using averaging. Figure 5 shows the concatenation of splits and different layers to blend different splits.

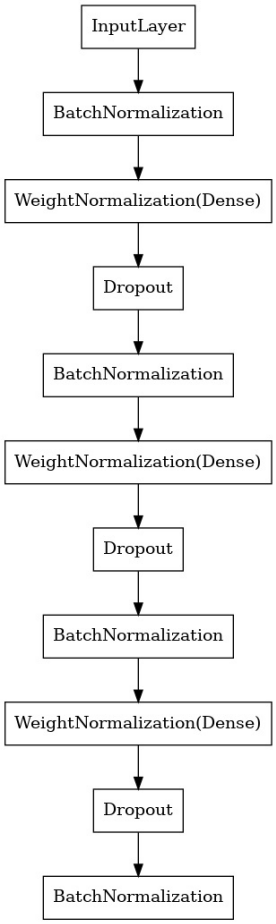

Figure 4. Architecture of the network (A single sample split and its layers)

The data split among train and test samples has not been entirely randomized from the analysis. The signatures in the training data at low/high doses and different time points. However, the data is not split by drug. Some drugs appeared only in training, while some in training and testing samples. From the dataset analysis, the dominant seven drugs contributing to 10 targets from a total of 206 targets (Table 2) are presented in Table 3. For each of the ten targets, the drugs contribute over 50% of the sig-ids. Initially, this may cause biases in prediction.

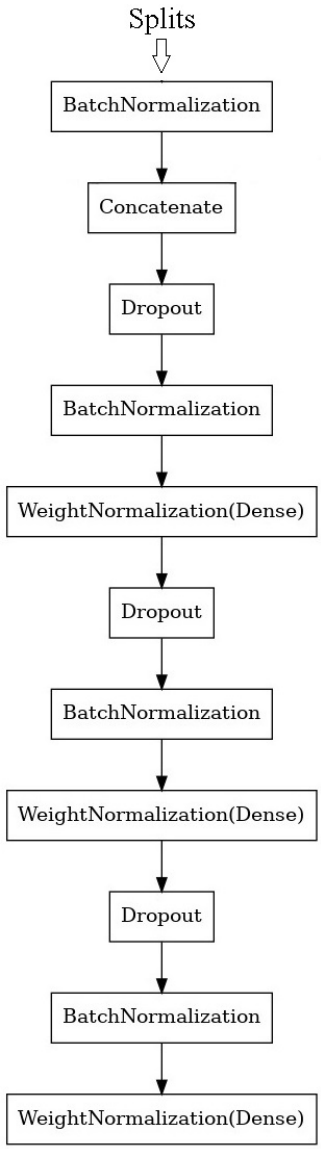

Figure 5. Architecture of the network (concatenation of the different splits and the layers)

Table 2. List of all targets involved in the MoA dataset

Sig_id	Sig_id
5-alpha_reductase_inhibitor	hdac_inhibitor
11-beta-hsd1_inhibitor	histamine_receptor_agonist
acat_inhibitor	histamine_receptor_antagonist
acetylcholine_receptor_agonist	histone_lysine_demethylase_inhibitor
acetylcholine_receptor_antagonist	histone_lysine_methyltransferase_inhibitor
acetylcholinesterase_inhibitor	hiv_inhibitor
adenosine_receptor_agonist	hmgcr_inhibitor
adenosine_receptor_antagonist	hsp_inhibitor
adenylyl_cyclase_activator	igf-1_inhibitor
adrenergic_receptor_agonist	ikk_inhibitor
adrenergic_receptor_antagonist	imidazoline_receptor_agonist
akt_inhibitor	immunosuppressant
aldehyde_dehydrogenase_inhibitor	insulin_secretagogue
alk_inhibitor	insulin_sensitizer
ampk_activator	integrin_inhibitor

Analgesic
androgen_receptor_agonist
androgen_receptor_antagonist
anesthetic_-_local
angiogenesis_inhibitor
angiotensin_receptor_antagonist
anti-inflammatory
antiarrhythmic
Antibiotic
anticonvulsant
Antifungal
antihistamine
Antimalarial
Antioxidant
antiprotozoal
Antiviral
apoptosis_stimulant
aromatase_inhibitor
atm_kinase_inhibitor
atp-sensitive_potassium_channel_antagonist
atp_synthase_inhibitor
atpase_inhibitor
atr_kinase_inhibitor
aurora_kinase_inhibitor
autotaxin_inhibitor
bacterial_30s_ribosomal_subunit_inhibitor
bacterial_50s_ribosomal_subunit_inhibitor
bacterial_antifolate
bacterial_cell_wall_synthesis_inhibitor
bacterial_dna_gyrase_inhibitor
bacterial_dna_inhibitor
bacterial_membrane_integrity_inhibitor
bcl_inhibitor
bcr-abl_inhibitor
benzodiazepine_receptor_agonist
beta_amyloid_inhibitor
bromodomain_inhibitor
btk_inhibitor
calcineurin_inhibitor
calcium_channel_blocker
cannabinoid_receptor_agonist
cannabinoid_receptor_antagonist
carbonic_anhydrase_inhibitor
casein_kinase_inhibitor
caspase_activator
catechol_o_methyltransferase_inhibitor
cc_chemokine_receptor_antagonist
cck_receptor_antagonist
cdk_inhibitor
chelating_agent
chk_inhibitor
chloride_channel_blocker
cholesterol_inhibitor
cholinergic_receptor_antagonist
coagulation_factor_inhibitor
corticosteroid_agonist
cyclooxygenase_inhibitor
cytochrome_p450_inhibitor
dihydrofolate_reductase_inhibitor
dipeptidyl_peptidase_inhibitor
diuretic
dna_alkylating_agent
dna_inhibitor
dopamine_receptor_agonist
dopamine_receptor_antagonist
egfr_inhibitor
elastase_inhibitor
erbb2_inhibitor
estrogen_receptor_agonist
estrogen_receptor_antagonist
faah_inhibitor
jak_inhibitor
kit_inhibitor
laxative
leukotriene_inhibitor
leukotriene_receptor_antagonist
lipase_inhibitor
lipoxygenase_inhibitor
lxr_agonist
mdm_inhibitor
mek_inhibitor
membrane_integrity_inhibitor
mineralocorticoid_receptor_antagonist
monoacylglycerol_lipase_inhibitor
monoamine_oxidase_inhibitor
monopolar_spindle_1_kinase_inhibitor
mtor_inhibitor
mucolytic_agent
neuropeptide_receptor_antagonist
nfkb_inhibitor
nicotinic_receptor_agonist
nitric_oxide_donor
nitric_oxide_production_inhibitor
nitric_oxide_synthase_inhibitor
norepinephrine_reuptake_inhibitor
nrf2_activator
opioid_receptor_agonist
opioid_receptor_antagonist
orexin_receptor_antagonist
p38_mapk_inhibitor
p-glycoprotein_inhibitor
parp_inhibitor
pdgfr_inhibitor
pdk_inhibitor
phosphodiesterase_inhibitor
phospholipase_inhibitor
pi3k_inhibitor
pkc_inhibitor
potassium_channel_activator
potassium_channel_antagonist
ppar_receptor_agonist
ppar_receptor_antagonist
progesterone_receptor_agonist
progesterone_receptor_antagonist
prostaglandin_inhibitor
prostanoid_receptor_antagonist
proteasome_inhibitor
protein_kinase_inhibitor
protein_phosphatase_inhibitor
protein_synthesis_inhibitor
protein_tyrosine_kinase_inhibitor
radiopaque_medium
raf_inhibitor
ras_gtpase_inhibitor
retinoid_receptor_agonist
retinoid_receptor_antagonist
rho_associated_kinase_inhibitor
ribonucleoside_reductase_inhibitor
rna_polymerase_inhibitor
serotonin_receptor_agonist
serotonin_receptor_antagonist
serotonin_reuptake_inhibitor
sigma_receptor_agonist
sigma_receptor_antagonist
smoothened_receptor_antagonist
sodium_channel_inhibitor
sphingosine_receptor_agonist
src_inhibitor
steroid
syk_inhibitor
tachykinin_antagonist
tgf-beta_receptor_inhibitor

farnesyltransferase_inhibitor	thrombin_inhibitor
fatty_acid_receptor_agonist	thymidylate_synthase_inhibitor
fgfr_inhibitor	tlr_agonist
flt3_inhibitor	tlr_antagonist
focal_adhesion_kinase_inhibitor	tnf_inhibitor
free_radical_scavenger	topoisomerase_inhibitor
fungal_squalene_epoxidase_inhibitor	transient_receptor_potential_channel_antagonist
gaba_receptor_agonist	tropomyosin_receptor_kinase_inhibitor
gaba_receptor_antagonist	trpv_agonist
gamma_secretase_inhibitor	trpv_antagonist
glucocorticoid_receptor_agonist	tubulin_inhibitor
glutamate_inhibitor	tyrosine_kinase_inhibitor
glutamate_receptor_agonist	ubiquitin_specific_protease_inhibitor
glutamate_receptor_antagonist	vegfr_inhibitor
gonadotropin_receptor_agonist	vitamin_b
gsk_inhibitor	vitamin_d_receptor_agonist
hcv_inhibitor	wnt_inhibitor

Table 3. Output of 10 targets and 7 unique over-represented drugs in the MoA dataset

Target	Drugs
Cdk inhibitor	24
Egfr inhibitor	25
flt3 inhibitor	17
Hmgcr inhibitor	7
Kit inhibitor	17
Nfkb inhibitor	19
Pdgfr inhibitor	21
Proteasome inhibitor	4
Raf inhibitor	9
Tubulin inhibitor	22

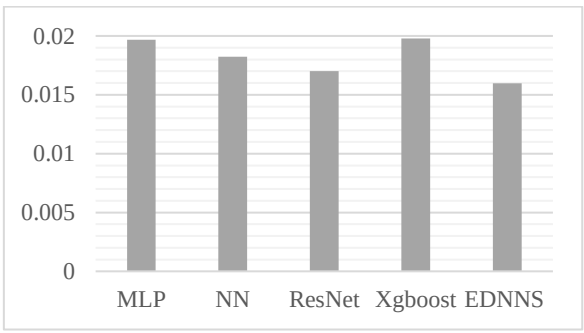

Figure 6. The log-loss-based comparison of the EDNNS with the MLP [37], Deep NN [38], ResNet [39], Xgboost [40]

For performance evaluation, the log losses are computed on the validation set. Figure 6 shows the comparison of the EDNNS with the MLP [37], Deep NN [38], ResNet [39], Xgboost [40]. MLP, Deep NN, ResNet, Xgboost models, and the EDNNS losses are 0.01968, 0.01823, 0.01701, 0.01978, and 0.01599, respectively. From the evaluation in the figure based on the log loss, the proposed ensemble EDNNS is more robust than the MLP, NN, ResNet, and Xgboost, having less log loss. The comparatively less loss thus means a more robust and accurate model for prediction. Generally, in biomedical and genomics data, the distribution is strongly affected by the type of sample, instrumentation inconsistencies, quality of reagent deviations, etc. The current study also argues that the biological data in raw form is not comparable due to systematic errors, which refer to the same data and the same person's experimental setup. Also, significantly different features stay smaller, thus assuming that the distributions of most samples in MoA are similar. Without prior experience with MoA or similar domains, the current study experimented with a concatenation of ML/DL framework to achieve better performances on MoA prediction. If the ML models learn the MoA distribution robustly with minimum loss and high accuracy, the obtained model can then predict a compound's MoA based on a specific cellular signature.

7. CONCLUSION

The current paper proposes NN based EDNNS approach in predicting multiple targets of the MoA responses of different samples. As such, the current study explored the answer to the question of the prediction of MoA based on gene expression and cell viability data using the ML paradigm. The current study evaluated MLP, Deep NN, ResNet, Xgboost and found that the proposed ensemble EDNNS is more robust than the MLP, Deep NN, ResNet, Xgboost having less loss. The comparatively less loss thus means a more robust and accurate model for prediction. This work can benefit the advanced drug discovery cause-effect by providing valuable insights and exciting directions for future research.

ACKNOWLEDGMENT

The Researcher would like to thank S. A. A. Mohammed.

REFERENCES

[1] Zhou, G., Myers, R., Li, Y., Chen, Y., Shen, X., Fenyk-Melody, J., Wu, M., Ventre, J., Doebber, T., Fujii, N., Musi, N., Hirshman, M.F., Goodyear, L.J., Moller, D.E. (2001). Role of AMP-activated protein kinase in mechanism of metformin action. The Journal of Clinical Investigation, 108: 1167-74. https://doi.org/10.1172/jci13505

[2] Wu, J., Li, Q., Bezprozvanny, I. (2008). Evaluation of Dimebon in cellular model of Huntington's disease. Molecular Neurodegeneration, 3: 1-11. https://doi.org/10.1186/1750-1326-3-15

[3] Dumont, J.E., Dremier, S., Pirson, I., Maenhaut, C. (2002). Cross signaling, cell specificity, and physiology. American Journal of Physiology-Cell Physiology, 283: C2-C28. https://doi.org/10.1152/ajpcell.00581.2001

[4] Mervin, L.H., Afzal, A.M., Drakakis, G., Lewis, R., Engkvist, O., Bender, A. (2015). Target prediction utilising negative bioactivity data covering large

chemical space. Journal of Cheminformatics, 7: 1-16. https://doi.org/10.1186/s13321-015-0098-y

[5] Bickle, M. (2010). The beautiful cell: High-content screening in drug discovery. Analytical and Bioanalytical Chemistry, 398: 219-226. https://doi.org/10.1007/s00216-010-3788-3

[6] Bray, M.A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C., Carpenter, A.E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nature Protocols, 11(9): 1757-1774. https://doi.org/10.1038/nprot.2016.105

[7] Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., Golland, P., Sabatini, D.M. (2006). CellProfiler: Image analysis software for identifying and quantifying cell phenotypes. Genome Biology, 7: 1-11. https://doi.org/10.1186/gb-2006-7-10-r100

[8] Held, M., Schmitz, M.H.A., Fischer, B., Walter, T., Neumann, B., Olma, M.H., Peter, M., Ellenberg, J., Gerlich, D.W. (2010). CellCognition: Time-resolved phenotype annotation in high-throughput live cell imaging. Nature Methods, 7: 747-754. https://doi.org/10.1038/nmeth.1486

[9] Rajaram, S., Pavie, B., Wu, L.F., Altschuler, S.J. (2012). PhenoRipper: Software for rapidly profiling microscopy images. Nature Methods, 9: 635-637. https://doi.org/10.1038/nmeth.2097

[10] Chandrasekaran, S.N., Ceulemans, H., Boyd, J.D., Carpenter, A.E. (2020). Image-based profiling for drug discovery: Due for a machine-learning upgrade? Nature Reviews Drug Discovery, 20: 145-159. https://doi.org/10.1038/s41573-020-00117-w

[11] Wollmann, T., Gunkel, M., Chung, I., Erfle, H., Rippe, K., Rohr, K. (2019). GRUU-Net: Integrated convolutional and gated recurrent neural network for cell segmentation. Medical Image Analysis, 56: 68-79. https://doi.org/10.1016/j.media.2019.04.011

[12] Caicedo, J.C., Goodman, A., Karhohs, K.W., Cimini, B.A., Ackerman, J., Haghighi, M., Heng, C., Becker, Y., Doan, M., McQuin, C., Rohban, M., Singh, S., Carpenter, A.E. (2019). Nucleus segmentation across imaging experiments: The 2018 data science bowl. Nature Methods, 16: 1247-1253. https://doi.org/10.1038/s41592-019-0612-7

[13] Saei, A.A., Gullberg, H., Sabatier, P., Beusch, C.M., Johansson, K., Lundgren, B., Arvidsson, P.I., Arnér, E.S.J., Zubarev, R.A. (2020). Comprehensive chemical proteomics for target deconvolution of the redox active drug auranofin. Redox Biology, 32: 101491. https://doi.org/10.1016/j.redox.2020.101491

[14] Medo, M., Aebersold, D.M., Medová, M. (2019). ProtRank: Bypassing the imputation of missing values in differential expression analysis of proteomic data. BMC Bioinformatics, 20: 1-12. https://doi.org/10.1186/s12859-019-3144-3

[15] Muelas, M.W., Roberts, I., Mughal, F., O'Hagan, S., Day, P.J., Kell, D.B. (2020). An untargeted metabolomics strategy to measure differences in metabolite uptake and excretion by mammalian cell lines. Metabolomics, 16: 1-12. https://doi.org/10.1007/s11306-020-01725-8

[16] Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., Sullivan, M. (2011). PhosphoSitePlus: A comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Research, 40: D261-D270. https://doi.org/10.1093/nar/gkr1122

[17] Brückner, A., Polge, C., Lentze, N., Auerbach, D., Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. International Journal of Molecular Sciences, 10: 2763-2788. https://doi.org/10.3390/ijms10062763

[18] Tian, B., Zhao, C., Gu, F., He, Z. (2017). A two-step framework for inferring direct protein-protein interaction network from AP-MS data. BMC Systems Biology, 11: 17-25. https://doi.org/10.1186/s12918-017-0452-y.

[19] Huang, H., Jedynak, B., Bader, J.S. (2005). Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. PLoS Computational Biology, Preprint: 3(11): e214. https://doi.org/10.1371/journal.pcbi.0030214.eor

[20] Zhu, X., Gerstein, M., Snyder, M. (2007). Getting connected: Analysis and principles of biological networks. Genes & Development, 21: 1010-1024. https://doi.org/10.1101/gad.1528707

[21] Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., et al. (2020). A reference map of the human binary protein interactome. Nature, 580: 402-408. https://doi.org/10.1038/s41586-020-2188-x

[22] Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T. (2018). The human transcription factors. Cell, 172: 650-665. https://doi.org/10.1016/j.cell.2018.01.029

[23] Park, P.J. (2009). ChIP: Advantages and challenges of a maturing technology. Nature Reviews Genetics, 10: 669-680. https://doi.org/10.1038/nrg2641

[24] von Mering, C. (2004). STRING: Known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Research, 33: D433-D437. https://doi.org/10.1093/nar/gki005

[25] Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., Palsson, B.Ø. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences, 104: 1777-1782. https://doi.org/10.1073/pnas.0610772104

[26] Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Research, 29(8): 1363-1375. https://doi.org/10.1101/337915

[27] Türei, D., Korcsmáros, T., Saez-Rodriguez, J. (2016). OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. Nature Methods, 13: 966-967. https://doi.org/10.1038/nmeth.4077

[28] Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Science, 30: 187200. https://doi.org/10.1002/pro.3978

[29] Domingo-Fernández, D., Mubeen, S., Marín-Llaó, J.,

Hoyt, C.T., Hofmann-Apitius, M. (2019). PathMe: Merging and exploring mechanistic pathway knowledge. BMC Bioinformatics, 20: 1-12. https://doi.org/10.1186/s12859-019-2863-9

[30] Bradley, G., Barrett, S.J. (2017). CausalR: Extracting mechanistic sense from genome scale data. Bioinformatics, 33: 36703672. https://doi.org/10.1093/bioinformatics/btx425

[31] Wiwie, C., Baumbach, J., Röttger, R. (2015). Comparing the performance of biomedical clustering methods. Nature Methods, 12: 10331038. https://doi.org/10.1038/nmeth.3583

[32] Karim, M.R., Beyan, O., Zappa, A., Costa, I.G., Rebholz-Schuhmann, D., Cochez, M., Decker, S. (2020). Deep learning-based clustering approaches for bioinformatics. Briefings in Bioinformatics, 22: 393-415. https://doi.org/10.1093/bib/bbz170

[33] Khan, S.A., Virtanen, S., Kallioniemi, O.P., Wennerberg, K., Poso, A., Kaski, S. (2014). Identification of structural features in chemicals associated with cancer drug response: A systematic data-driven analysis. Bioinformatics, 30: i497-i504. https://doi.org/10.1093/bioinformatics/btu456

[34] Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O. (2018). Multi-Omics Factor Analysis: A framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology, 14(6): e8124. https://doi.org/10.15252/msb.20178124

[35] Bender, A., Scheiber, J., Glick, M., Davies, J.W., Azzaoui, K., Hamon, J., Urban, L., Whitebread, S., Jenkins, J.L. (2007). Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure. ChemMedChem: Chemistry Enabling Drug Discovery, 2(6): 861873. https://doi.org/10.1002/cmdc.200700026

[36] Mechanisms of action (MoA) prediction. Kaggle. https://www.kaggle.com/competitions/lish-MoA/data, accessed Apr. 17, 2022.

[37] Neural network models (supervised). Scikit. https://scikit-learn.org/stable/modules/neural_networks_supervised.html, accessed Apr. 17, 2022.

[38] Deep Learning Convolutional Neural Network. (2016). World Scientific, pp. 41-55. https://doi.org/10.1142/9789813146464_0005

[39] Li, Z., Lin, Y., Elofsson, A., Yao, Y. (2020). Protein contact map prediction based on ResNet and DenseNet. BioMed Research International, 2020: 1-12. https://doi.org/10.1155/2020/7584968

[40] Sagi, O., Rokach, L. (2021). Approximating XGBoost with an interpretable decision tree. Information Sciences (Ny), 572: 522-542. https://doi.org/10.1016/j.ins.2021.05.055