MERAMALNET: A DEEP LEARNING CONVOLUTIONAL NEURAL NETWORK FOR BIOACTIVITY PREDICTION IN STRUCTURE-BASED DRUG DISCOVERY

Hentabli Hamza[1], Naomie Salim[1], Maged Nasser[1], Faisal Saeed[2]

[1]Faculty of Computing, Universiti Teknologi Malaysia, Malaysia
[2]College of Computer Science and Engineering,
Taibah University, Medina, Saudi Arabia

ABSTRACT

According to the principle of similar property, structurally similar compounds exhibit very similar properties and, also, similar biological activities. Many researchers have applied this principle to discovering novel drugs, which has led to the emergence of the chemical structure-based activity prediction. Using this technology, it becomes easier to predict the activities of unknown compounds (target) by comparing the unknown target compounds with a group of already known chemical compounds. Thereafter, the researcher assigns the activities of the similar and known compounds to the target compounds. Various Machine Learning (ML) techniques have been used for predicting the activity of the compounds. In this study, the researchers have introduced a novel predictive system, i.e., MaramalNet, which is a convolutional neural network that enables the prediction of molecular bioactivities using a different molecular matrix representation. MaramalNet is a deep learning system which also incorporates the substructure information with regards to the molecule for predicting its activity. The researchers have investigated this novel convolutional network for determining its accuracy during the prediction of the activities for the unknown compounds. This approach was applied to a popular dataset and the performance of this system was compared with three other classical ML algorithms. All experiments indicated that MaramalNet was able to provide an interesting prediction rate (where the highly diverse dataset showed 88.01% accuracy, while a low diversity dataset showed 99% accuracy). Also, MaramalNet was seen to be very effective for the homogeneous datasets but showed a lower performance in the case of the structurally heterogeneous datasets.

KEYWORDS

Bioactive Molecules, Activity prediction model, Convolutional neural network, Deep Learning, biological activities

1. INTRODUCTION

The biological systems function via the physical interactions occurring between the molecules. Hence, it is important to determine the molecular binding for understanding the biological system and discovering novel drugs [1]. The pharmaceutical industries have devoted a lot of effort towards discovering novel drugs. This discovery could improve our quality of life, however, could also lead to many adverse effects [2], [3]. Hence, the pharmaceutical companies must ensure drug safety during the research stage, as the observation of adverse effects during late clinical phases could lead to heavy financial losses. However, despite the development of

Natarajan Meghanathan et al. (Eds): SIPP, BIGML, DaKM, SOEN, AISC - 2020
pp. 21-37, 2020. © CS & IT-CSCP 2020 DOI: 10.5121/csit.2020.100203

computational systems for the past 30 years, they are inaccurate while predicting the molecular binding, and, physical experiments have to be conducted for determining the binding [3], [4].
The accurate molecular binding prediction could decrease the time required for discovering novel treatments, eliminating the toxic molecules in the initial developmental stages and for guiding studies towards medicinal chemistry [5]. Despite the requirement of powerful, but, versatile tools for determining the side effects of the novel drugs, none have been discovered till date. This problem can be solved by implementing computational models which have been obtained using the standard Quantitative Structure-Activity Relationships (QSAR) [6], [7].

In the similarity searching strategy, the activities of unknown compounds (target) are predicted by comparing them with the known chemical compounds. Thereafter, the researcher assigns the activities of similar compounds to the target compounds. Though many of the target prediction techniques have been successful, some problems still exist. Researchers have applied different techniques for predicting different target subsets for the same molecule [8]–[10]. One study [11] used the Multilevel Neighbourhoods of Atoms (MNA) structural descriptor system for activity prediction. MNA of the molecule is generated by the connection table and the table of atoms, representing every compound. Every descriptor possessed a specific integer number based on its dictionary. The molecular similarity was based on the Tanimoto coefficient, and, the compound activities were predicted using the activities of the most similar known compounds.

The popular ML algorithms, using the compound classification method for activity prediction (target), were Binary Kernel Discrimination (BKD) [12], Bayesian inference network for ligand-based virtual Screening [13], Naïve Bayesian Classifier (NBC) [14], Artificial Neural Networks (ANNs) [15] and Support Vector Machines (SVM) [16]. The Bayesian belief network classifier was used for predicting the ligand-based targets and their activities [1]. Here, the researchers introduced a novel approach, MaramalNet (Maramal means ‘predicting’ in Malay), which is a convolutional neural network that predicts the molecular bioactivity using a novel molecular matrix representation. Also, it is a deep learning system which incorporates the molecule’s substructural information for activity prediction.

2. DEFINITION AND RELATED WORK

2.1 Deep Learning

Deep learning is seen to dramatically improve the advanced artificial intelligent tasks such as speech recognition, object detection and machine translation [17]. The deep architectural nature of this technique is useful for solving the complex artificial intelligence-related problems [18]. Hence, researchers have used this technique in modern domains for several tasks like face recognition and object detection. This method has also been applied to many language models. For instance, [17] applied the recurrent neural networks for denoising the speech signals, [19] used the stacked autoencoders for determining the cluster pattern during gene expression. In another study,[20], the researchers used a neural model for generating images having differing styles. Also, [21] used the deep earning technology for a simultaneous analysis of sentiments from the multiple modalities.

The deep learning technology has undergone massive developments during the past few years. Empirical results showed that this technique was better than the other ML algorithms. This could due to the fact that this technique mimics the brain functioning and stacks multiple neural network layers one after another, like the brain model. According to [22], the Deep Learning machines show a better performance than the conventional ML tools as they also include the feature extraction method. However, till date, there exists no theoretical background for the deep learning technology. The deep learning techniques learn the feature hierarchies by using features from the higher hierarchical levels formed by the arrangement of the low-level features. The

learning features present at various abstraction levels allow the system to learn the complex functions which map the input and the resultant output from the data without depending on the human-developed features [22]. In the case of the image recognition systems, the conventional setup extracts the handcrafted features and feeds them to the SVM. However, the deep learning technology shows a better performance as it also optimises all the extracted features.

The biggest difference between the ML and deep learning technologies is their performance variations when the data volume increases. For a smaller dataset, the deep learning method performs inefficiently as it needs a huge data volume for proper understanding [21].

2.2 Convolutional Neural Network

The Convolutional Neural Network (CNN) is a type of deep feed-forward network which can be easily trained and generalised as compared to other networks having connectivity between the adjacent layers [23], [24]. CNN has been successfully used when other neural networks were unpopular, and currently, has been used in the computer vision community.

CNNs are designed for processing data which is in the form of multiple arrays, for instance, a grey-scale image made of 3×2D arrays with different pixel intensities. Various data modalities are presented as multiple arrays, like 1D for sequences and signals, including language; 2D for the audio or image spectrograms; and 3D for the volumetric or video images. The 4 major ideas which enable the CNNs to use the properties of the natural signals are shared weights, local connections, pooling and use of multiple layers [23], [24].

A classic CNN architecture (Figure 1) includes many stages. The initial stages are made of 2 types of layers: i.e., convolutional and pooling layers. The units within the convolutional layer can be organised in the feature maps, wherein every unit is linked to the local patches of the feature maps from the earlier layers through weights known as the filter bank. The output of the local weighted sum is passed through the non-linearity like the ReLU [25]. All the units within the feature map are seen to share one filter bank. The various feature maps within the layer use differing filter banks. This architecture is so composed to serve 2 purposes. Initially, in the case of array data like images, the local group of values are seen to be highly correlated and form distinctive and easily detectable local motifs. Secondly, the local statistics of the images or other signals are seen to be invariant to the location. Hence, if the motif is seen within one section of the image, it can also be present elsewhere. Thus, this network relies on the fact that the units at the different locations share the same weights and can be detected using the similar pattern from the other parts in the array. Mathematically, discrete convolution is the main filtering operation which is applied in the feature maps; hence, it is so named.

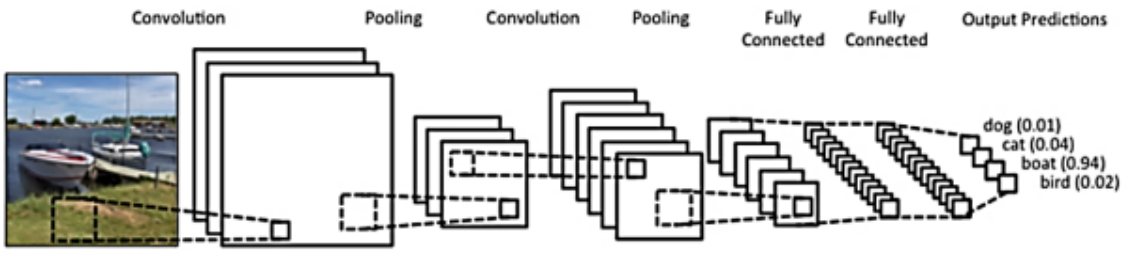


Fig. 1. Architecture of the CNN for Image Classification.

While the convolutional layer detects the local combination of the features based on the earlier layer, the pooling layer merges the semantically similar features into a single feature. Due to the relative position of these features, the motif formation can vary and the reliable detection of the motif is carried out by the coarse-graining of its position in every feature. The general pooling unit can compute a maximal number of the local patch of units into a single feature map.

As described in Figure 2, for the image classification, the CNN technique detects the edges from the raw pixels in Layer 1, and thereafter, uses the edges for detecting the simple shapes in Layer 2. Then, it uses these shapes for detecting the simpler shapes within the Layer 2 and also uses these shapes for determining the high level features, like the face shape in the higher layers. The final layer is the classifier which uses such high-level features [26].

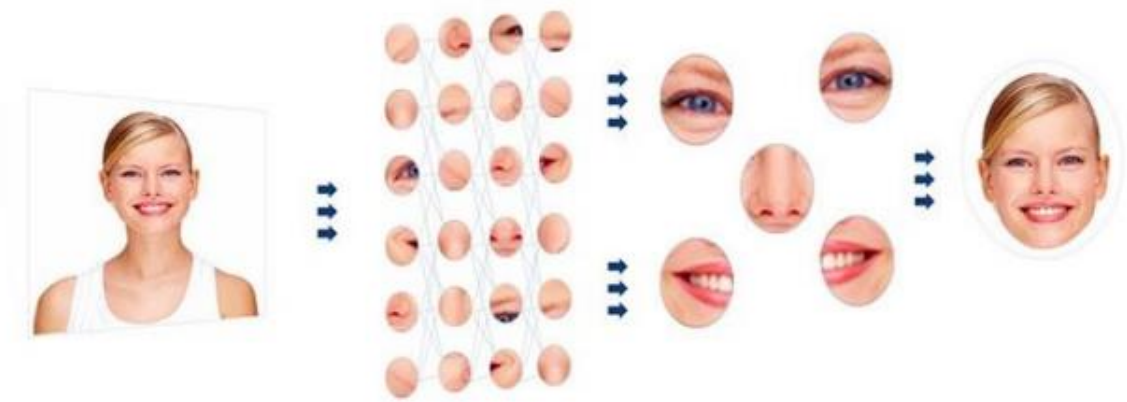


Fig. 2. Eyeris' Deep Learning-based facial feature extraction method based on CNN[26].

2.3 Convolutional Neural Network for the Prediction of the Biological Activities

In [27], the architecture of the Merck Molecular Activity Kaggle Challenge, based on the Multi-Task Deep Neural Network (MT-DNN) [28] showed the best performance. This architecture could train the neural network using multiple output neurons, wherein every neuron predicts the input molecule's activity using different assays. Also, [29]–[31] showed that MT-DNN could be scaled to include large biochemical databases like PubChem Bioassays [32] and ChEMBL [33].

However, there are many limitations associated with the ligand-based processes, like the MT-DNN. Firstly, these techniques are limited to those targets having a lot of prior available data, and hence, they are unable to make predictions for the novel targets. Secondly, the current deep neural networks designed for the ligand-based models also use some molecular fingerprints, like ECFP [34], as their input data. This type of input encoding restricts the feature discovery to the composition of the specific molecular structures that are defined by the fingerprinting procedure [29], [35], which eliminates its capacity to discover the arbitrary features. Thirdly, as these models are blind towards the target, they cannot elucidate the potential molecular interactions.

Another popular strategy used for library designing includes the application of the similarity principle [36], where the structurally similar compounds exhibit similar biological properties. But, researchers [37] have shown that such an empirical guideline is often unsuccessful, as the minor structural modifications could diminish the pharmacological activities of the ligand which is used for describing the molecular similarity within the substructures.

For addressing these limitations, the researchers in this study have proposed a novel matrix representation for the chemical compounds, i.e., mol2matrix, which is based on the molecular substructural similarities with a set of other molecules. Thereafter, the similarity values are

determined and the molecules arranged in the matrix. This technique can be used for many deep learning applications like prediction, virtual screening, molecular classification and molecular search. The following sections describe the design and the development of the MaramalNet approach. Also, the researchers have assessed its performance level by conducting several complex experiments which were based on the structure and bioactivity prediction.

3. MATERIALS AND METHODS

Initially, the researchers have described the construction of various experimental benchmarks used for testing the system. Thereafter, they have described the data encoding and Input representation system along with the design of the deep convolutional network.

3.1. Data Sets

Experiments were conducted over the most popular cheminformatics database: the MDL Drug Data Report (MDDR) [38]–[40] which has been used in our previous studies [1], [41]–[44]. This database consisted of 8294 molecules and contains 11 activity classes, which involve structurally homogeneous and heterogeneous actives, as shown in Table 1. Each row in the tables contains an activity class, the number of molecules belonging to the class, and the diversity of the class, which was computed as the mean pairwise Tanimoto similarity calculated across all pairs of molecules in the class with the ECFP4 (extended connectivity).

Table 1. MDDR Activity Classes Data Set

Activity Index	Activity class	Active molecules	Pairwise Similarity
31420	renin inhibitors	1130	0.290
71523	HIV protease inhibitors	750	0.198
37110	thrombin inhibitors	803	0.180
31432	angiotensin II AT1 antagonists	943	0.229
42731	substance P antagonists	1246	0.149
06233	substance P antagonists	752	0.140
06245	5HT reuptake inhibitors	359	0.122
07701	D2 antagonists	395	0.138
06235	5HT1A agonists	827	0.133
78374	protein kinase C inhibitors	453	0.120
78331	cyclooxygenase inhibitors	636	0.108

3.2. Input Representation

The feature extraction step is very important for analysing the data in the ML and NLP processes. This step helps in determining the interpretable data representation for the machines which could improve the performance of these learning algorithms. The application of inappropriate features could decrease the performance of even the best algorithms, whereas simple techniques perform very well if appropriate features are applied. The feature extraction is carried out unsupervised or even manually. In this study, the researchers have proposed an unsupervised distributed representation of the various chemical compounds.

A novel method was proposed called the mol2matrix (molecule to matrix). This technique could be used in cheminformatics for many problems related to the virtual screening, classification, biological activity prediction, similarity measurements and a substructure search of the

molecules. Here, every compound was embedded within an $n \times n$ matrix, which characterised the various molecular properties.

The distributed representation was seen to be a successful and popular ML approach [46], [47]. This approach involved encoding and storage of information within the system by interacting with the other compounds. The distributed representation technique was inspired by the human memory structure, wherein all memories are stored in a "content-addressable" manner. The content-based storage efficiently recalls all memories based on their partial description. Since these content-addressable thoughts and their properties are stored in a close proximity, the systems possess a viable infrastructure for generalising the features for any item.

The continuous vector representation, which acts like a distributed representation of words, was used in the Natural Language Processing (NLP) system for efficiently representing the semantic/syntactic units having multiple applications. In the model, every word was embedded with the vector in the *n-dimensional* space. The similar words had closer vectors, like "King, Queen" and "Woman, Man", wherein the similarity was based on the syntax and semantics. These vectors were trained based on the idea that the meaning behind the words was characterised by their context, i.e., neighbouring words. Hence, the various words along with their context were considered as the positive training samples [45]. They observed very interesting patterns by training the word vectors with the Skip-gram in the natural language. The words, having a similar vector representation, exhibit multiple similarity degrees. For example, Figure 3 shows that the words $\overrightarrow{King} - \overrightarrow{Man} + \overrightarrow{Woman}$ resemble their closest vector with the word $\overrightarrow{Queen}$ [46].

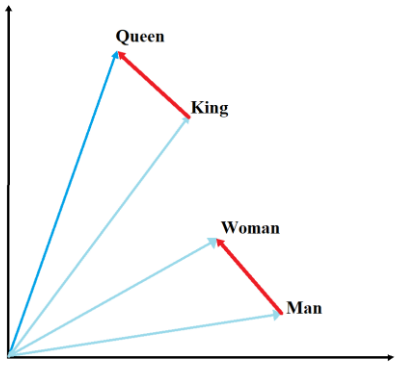


Fig. 3. Word2Vec wherein the words with similar vector representations display multiple similarity degrees.

Deep learning possesses the ability for constructing abstract features and this helps in predicting toxicity or biological activities. Here, the researchers have determined new chemical compound patterns for facilitating their biochemical and biophysical interpretation. Mol2matrix showed a similar molecular representation as the images represented by the Deep Learning technique.

The biological activity of a compound is an adverse property which affects its potential of becoming being marketed as a drug. The toxic or biological properties of a compound are based on their chemical structure, particularly, their substructures, which are identified as functional groups or toxicophores. Many toxicophores have been identified and described earlier [47]–[50].

In this study, the researchers predicted the biological activities using the molecules' distributed representation. This approach included the encoding and storage of information regarding the chemical compounds by establishing their interactions and similarities to the standard

toxicophores. With this in mind, the researchers assessed the similarities of every compound with the known 4096 toxicophore features, i.e., substructural patterns that represented the functional groups reported earlier [51].

The Kazius dataset (Figure 4) comprises of a group of 29 toxicophores, developed from the mutagenicity dataset after applying a novel toxicophore selection and validation criterion. It also consists of statistical, mechanistic and chemical information. These approved toxicophores are used for classifying and predicting the mutagenicity of various compounds in other datasets and display a high accuracy along with good sensitivity and specificity values. The researchers concluded that this set of toxicophores were very helpful in the biological activity prediction of any chemical compound [52]. The Kazius database consists of 4337 compounds, which are converted to ECFP4 by Pipeline Pilot. The first character in the name of the fingerprint, i.e., E, represents the atom abstraction process used for assigning the initial atom code that was based on the number of connections with an atom, the type of element, charge and the atomic mass fingerprints, and is thereafter folded to the final size of 1024 [53].

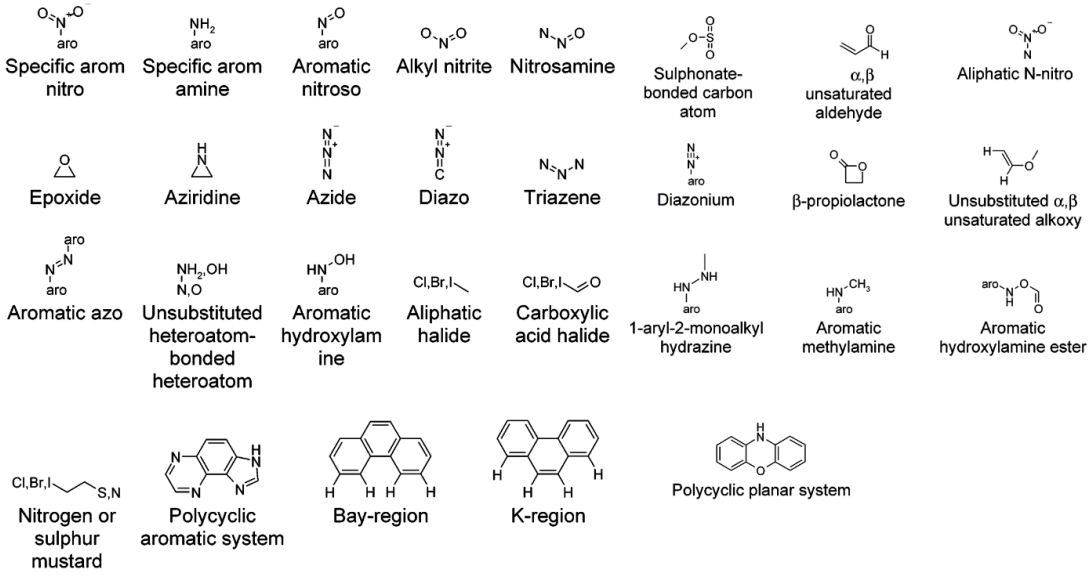


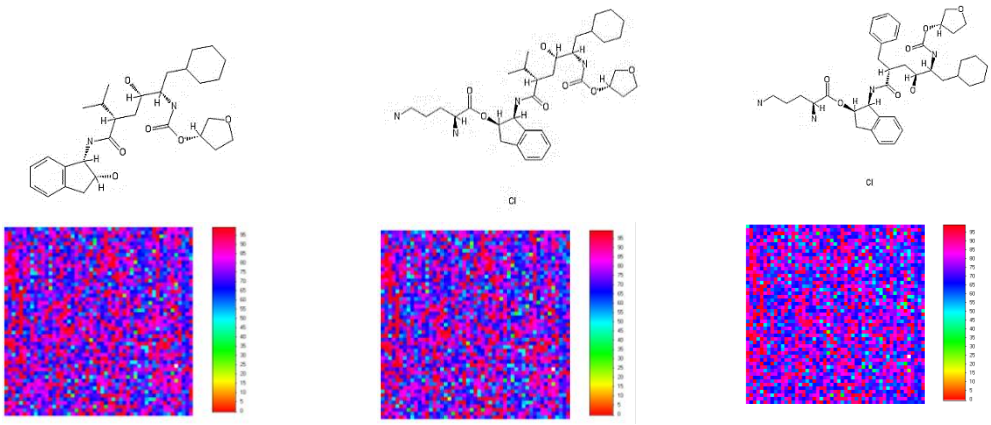
Fig. 4. A set of approved 29 toxicophores within the Kazius Dataset[51].

In this study, the researchers have proposed the mol2matrix for representing every molecule in the 64 × 64 matrix. This matrix comprises of the compounds displaying Tanimoto similarities to the 4096 toxicophore features presented in the Kazius dataset. After eliminating the 241 toxicophore features with the highest similarity, the Tanimoto-based Similarity Searching (TAN) [44] technique applied the binary form of the Tanimoto coefficient to the binary data. A similarity score, S_{xy}, was used for computing the similarities between the 2 molecular ECFP4 fingerprints, i.e., X and Y, with a length of 1024, wherein 'A' represented the number of bits present in the X and Y fingerprints, 'B' represented the number of bits that were present only in X, while 'C' represented the number of bits presents only in Y.

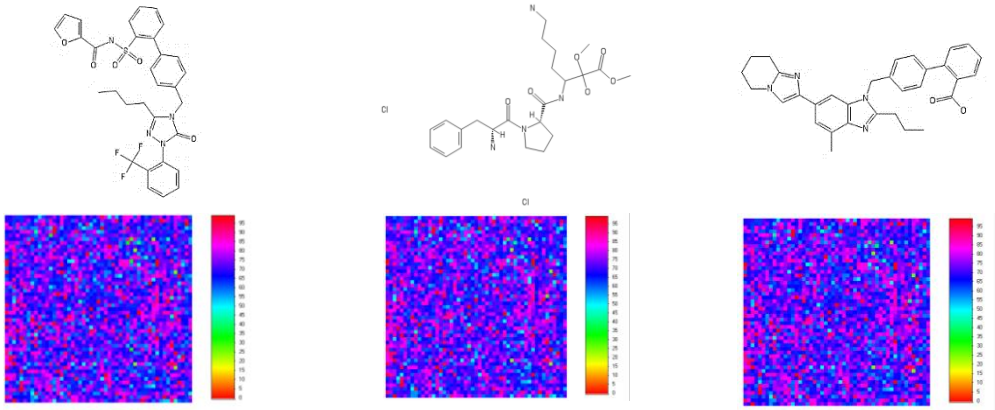
$$S_{xy} = \frac{A}{\mathrm{A} + \mathrm{B} + \mathrm{C}}$$

Every row in the matrix is filled with the molecules based on the order of their toxicophore features after comparing them to the Kazius dataset. This mol2matrix representation helps in visualising and characterising every molecule in the matrix based on their interactions and

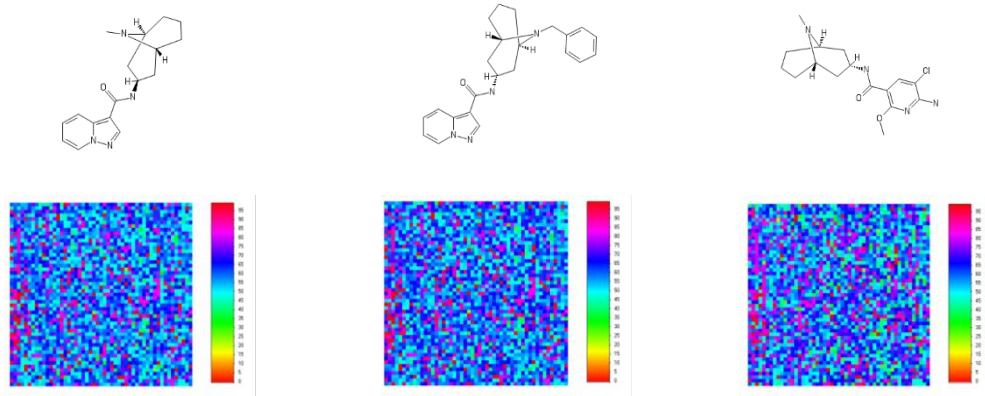
similarities with the functional groups. Thereafter, this matrix describes the toxic properties of the chemical compound. The mol2matrix is a good tool for describing the computational toxicology as it constructs the abstract chemical features. Figure 5 describes the molecules having different biological activities and classed in the MDDR dataset that were used in this study, along with their mol2matrix representation.



a)HIV protease inhibitors activity class



b) Angiotensin II AT1 antagonists activity class



c) 5HT3 antagonist activity class

Fig. 5. Examples describing 9 molecules that were categorised in 3 biological classes of the MDDR datasets and were used in this study along with their mol2matrix representation.

For studying the performance of the mol2matrix representation, the researchers also plotted the scatter graphs using the 8294 molecules which were categorised into 10 different biological

activity classes in the MDDR dataset (Figure 6). These scatter plots are used for determining the relationship between the different molecules within the same class, which was based on their individual representation that was reduced to a 3D structure using the Principal Component Analysis (PCA) technique for representing their features. As seen in the figure, the mol2matrix representation was not overlapping and could be observed easily and thereafter. Also, the biological activities of the molecules could be segregated. This shows that the proposed mol2matrix method can be successfully applied for the molecular representation and the biological activity prediction of different chemical compounds.

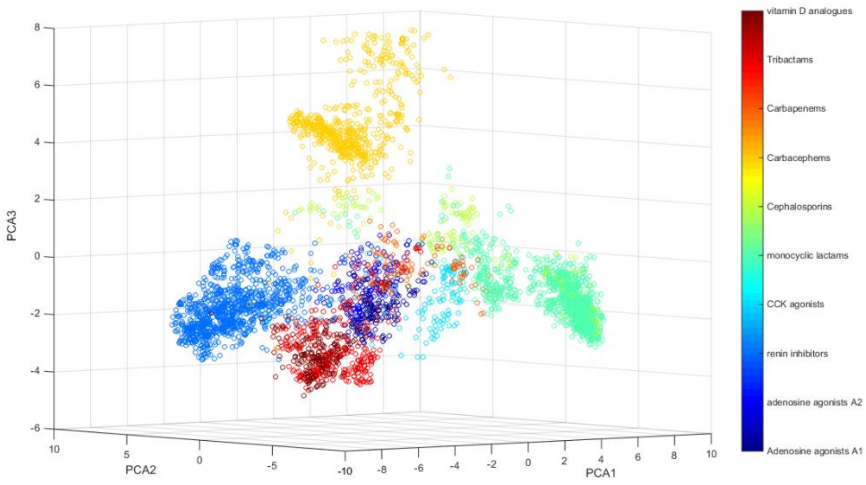


Fig. 6. 3D-scatter plots based on the mol2matrix representation of 8294 different molecules that were selected from the 10 biological activity classes of the MDDR dataset.

3.3. Network Architecture

After collecting all data, the researchers investigated the various model architectures. They considered the convolutional architecture with fully connected layers as the default architecture. Such architecture is appropriate for the multi- and high-dimensional data, like 2D images or genomic data, the researchers designed the MaramalNet layer configurations using the Krizhevsky principles[23] can view the source code through[54]. The configuration followed the generic design described earlier [23]. Fig.7 illustrates the proposed MaramalNetconfiguration.

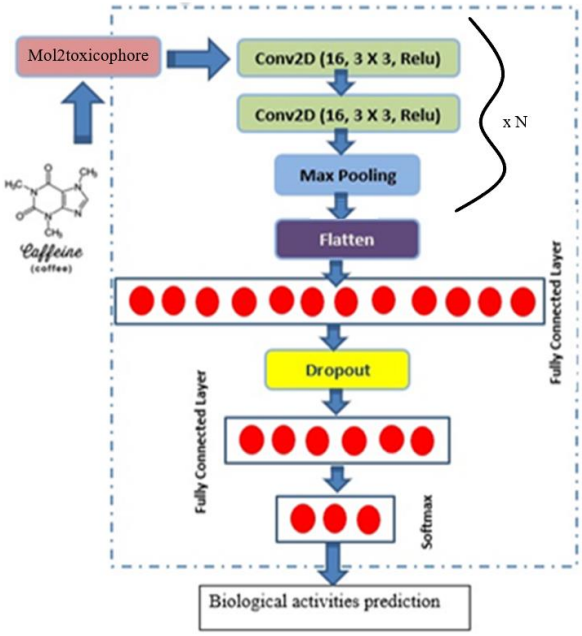


Figure 7 . The proposed MaramalNet configuration

Generally, the target prediction is carried out as follows:
The problem involves predicting if the given chemical compound, i, is active against the target, t. This information is encoded in the binary form, y_{it}, wherein $y_{it} = 1$, for an active compound, and is $y_{it} = 0$, if not. The problem also requires the compound behaviour prediction on *m* targets, simultaneously. During the training stage, a standard backpropagation algorithm is used for determining the CNN and minimising the cross-entropy of the targets and the output layer activation.

3.4. Machine Learning Algorithms

The researchers compared their proposed technique to 3 other available ML algorithms within the WEKA-Workbench [55], i.e., the Naive Bayesian classifier (NaiveB) [56], SVM classifier (known as the LibSVM, LSVM) [57] and a neural network classifier (RBFN) [58]. Determining the ideal classifier parameters is a very tiring process. However, the WEKA-Workbench helps in determining the best probable setup for the LSVM classifier. In this study, the LSVM has been applied to the linear kernel and the values of 0.1, 1.0, and 0.001 have been allocated to the Gamma, Cost, and Epsilon parameters, respectively. The researchers used a supervised discretisation technique for converting the numeric attributes to the nominal attributes in the NaiveB classifier and a minimal standard deviation limit of 0.01 was set in the RBFN classifier. All remaining parameters were kept default for every classifier in the WEKA-Workbench.

4. RESULTS AND DISCUSSION

The proposed code has been implemented in the Theano [59], which is a public deep learning software, based on the Keras [60]. The weights in the neural networks were initialised according to the Keras settings. All layers in the deep network were initialised simultaneously with the ADADELTA [61]. The complete network was trained using the Dell Precision T1700 CPU system with a 14GB memory and the professional-grade NVIDIA-Quadro discrete graphics. The deep network required 2 weeks for its training and testing.

4.1. Evaluation Measures

The researchers used a 10-fold cross-validation technique for validating all results of their proposed MaramalNet system. In this method, they divided the dataset into 10 sections, where 7 sections were used for training, while 3 were used for testing purposes. This procedure was repeated 10 times, hence, all compounds could be used within the test set at least once. Thus, every activity class could be tested against the other classes. Similar to other prediction techniques, the researchers determined the Area Under the receiver operating characteristic Curve (AUC) and used it as the quality criterion for assessing the performances of the various classification algorithms. AUC was estimated as follows:

$$\mathbf{AUC} = (\mathbf{sens} + \mathbf{spec})/\ 2 \quad (1)$$

Wherein sens and spec represent the sensitivity and specificity values, respectively, and are estimated as follows:

$$\mathbf{Sens} = \mathbf{tp}/\ (\mathbf{tp} + \mathbf{fn}) \quad (2)$$

$$\mathbf{Spec} = \mathbf{tn}/\ (\mathbf{tn} + \mathbf{fp}) \quad (3)$$

Wherein tp, tn, fp and fn are the no. of true positive, true negative, false positives, and false negatives, respectively. Where tp are the number of active molecules within the active set, while tn refers to the no. of inactive molecules that are selected in the inactive set. Meanwhile, fp and fn refer to the no. of active molecules present in the inactive set, and the no. of inactive molecules in the active set, respectively. In the model, a curve described the trade-off between the sensitivity and specificity, wherein sensitivity and specificity were defined as the efficiency of the model for identifying the positive and the negative labels, respectively. Furthermore, the Area Under the Curve (AUC) also assesses the model performance. When the AUC value of the prediction algorithm is nearer to 1, it is said to show a better performance.

4.2. Results

In this study, the researchers proposed MaramalNet, which was a novel ligand-based activity prediction or target fishing method for unknown chemical compounds. MaramalNet is a convolutional neural network, having a new molecular matrix representation, and is used for molecular bioactivity prediction. Furthermore, it is a deep learning system which incorporates the substructure information regarding the molecules for making predictions. Hence, the proposed MaramalNet technique was compared to 3 other ML algorithms present in the WEKA-Workbench, i.e., NaiveB, LSVM, and RBFN using optimal parameters.

Table 2 display the Sensitivity, Specificity and the AUC values for the MDDR dataset used in the study. Though a visual inspection of these tables could be used for comparing the prediction accuracy performance of the 4 algorithms, the researchers employed a quantitative technique of one-way ANOVA. This technique quantified the level of agreement observed between the multiple sets which ranked the same group of objects.

Table 2. Sensitivity, Specificity and AUC rates for the Prediction Models using the MDDR dataset.

activity index	MeramalNet			NaïveB			RBFN			LSVM		
	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC
0.99	1	0.99	1	1	1	0.96	0.96	0.96	1	1	1	0.99
0.97	0.99	0.98	1	1	1	0.95	0.97	0.96	1	1	1	0.97
0.95	1	0.98	1	0.99	1	0.44	1	0.72	0.98	1	0.99	0.95
0.96	1	0.98	1	1	1	0.8	1	0.9	1	1	1	0.96
0.98	1	0.99	0.94	1	0.97	0.43	1	0.72	0.99	1	1	0.98
0.92	1	0.96	0.94	1	0.97	0.53	1	0.76	0.98	1	0.99	0.92
0.92	0.99	0.95	0.84	0.99	0.91	0.78	0.97	0.87	0.96	0.99	0.98	0.92
0.96	1	0.98	0.82	0.99	0.91	0.75	0.97	0.86	0.94	1	0.97	0.96
0.96	0.99	0.98	0.88	0.97	0.92	0.66	0.98	0.82	0.96	0.99	0.98	0.96
0.94	1	0.97	0.65	0.99	0.82	0.74	0.96	0.85	0.91	1	0.95	0.94
0.93	0.98	0.95	0.82	0.94	0.88	0.59	0.96	0.78	0.94	0.98	0.96	0.93

In this study, the researchers have applied the one-way ANOVA technique for evaluating the performance of all the 4 algorithms. Hence, in this case, the MDDR activity classes that were described earlier in Tables 1, were considered to be judges, while the parameters of Sensitivity, Specificity and AUC, which were measured for the different prediction algorithms, were considered to be objects. This test showed an output in the form of the *p-value*, median and the variance. In Figure 8, the researchers have presented the results of the one-way ANOVA test after comparing the sensitivity values for the MeramalNet, NaiveB, RBFN and the LSVM algorithms. A very small *p-value* of 1.16×10^{-3} was observed which clearly indicated the high significance of difference between the algorithms. Furthermore, it could be seen that the MeramalNet algorithm displayed a good sensitivity value of 0.94. A larger variance was noted between the NaiveB and the LSVM ML algorithms, i.e., 0.15 and 0.23, respectively, in comparison to the MeramalNet algorithm. This highlights the diversity in the sensitivity values noted in the algorithms with a variance of 0.049. Meanwhile, those models exhibited an average sensitivity value of 0.90 and 0.74, respectively.

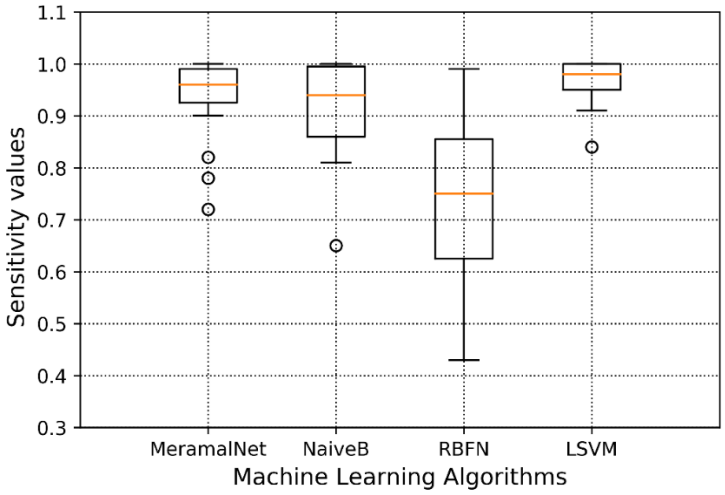


Fig. 8. Comparison of the sensitivity values for the MeramalNet, NaiveB, RBFN and LSVM algorithms using ANOVA

In Figure 9, the researchers have presented the results of the one-way ANOVA test after comparing the specificity values for the MeramalNet, NaiveB, RBFN and the LSVM algorithms. A larger variance was noted between the NaiveB and the RBFN ML algorithms, i.e., 0.01 and 0.04, respectively, in comparison to the MeramalNet algorithm. This highlights the diversity in the specificity noted in the algorithms with a variance of 0.0062. Also, the MeramalNet algorithm displayed a good specificity value of 1.0, while the NaiveB and RBFN algorithms showed a mean specificity value of 0.99 and 0.98, respectively. A small *p-value* of 6.25×10^{-5} was seen which indicated the significance of difference between the algorithms.

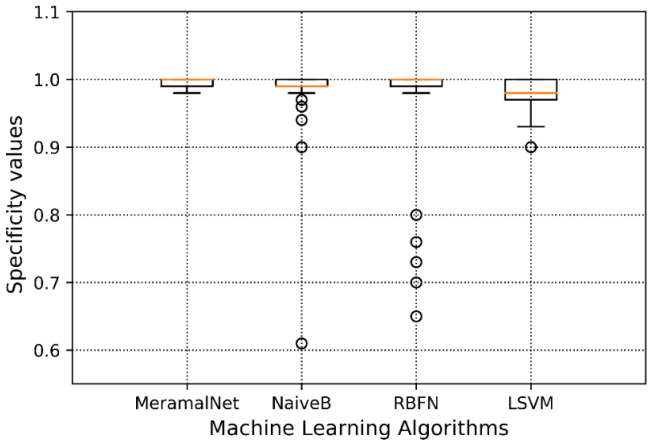


Fig. 9. Comparison of specificity values for the MeramalNet, NaiveB, RBFN and LSVM algorithms using ANOVA

In Figure 10, the researchers have presented the results of the one-way ANOVA test after comparing the AUC values for the MeramalNet, NaiveB, RBFN and the LSVM algorithms. A larger variance was noted between the LSVM, NaiveB and RBFN ML algorithms, i.e., 0.125, 0.083 and 0.033, respectively, in comparison to the MeramalNet algorithm. This highlights the diversity in the AUC values noted in the algorithms with a variance of 0.02. The MeramalNet algorithm displayed a good AUC value of 0.98, while the LSVM, NaiveB and RBFN algorithms showed a mean AUC value of 0.96, 0.99 and 0.85, respectively. A very small *p-value* of 1.6×10^{-14} was seen which indicated the significance of difference between the algorithms.

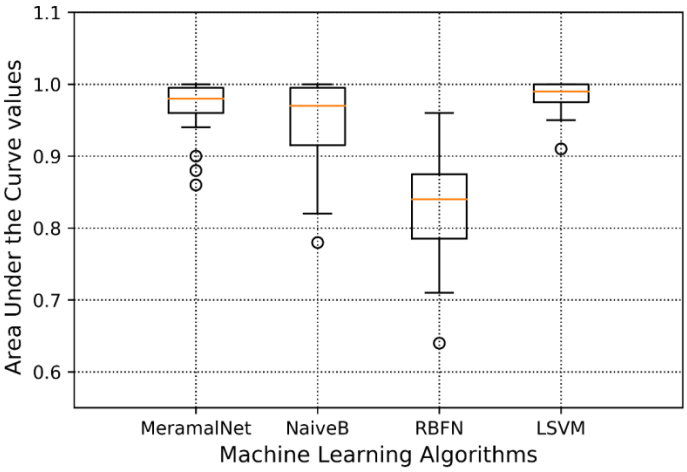


Fig. 10. Comparison of AUC values for the MeramalNet, NaiveB, RBFN and LSVM algorithms using ANOVA

A visual inspection of the one-way ANOVA results (Figures 16-18) indicated that the MaramalNet activity prediction technique was more applicable, convenient and exhibited less severe outliers compared to the NaiveB, RBFN and LSVM ML algorithms, thus, proving the efficacy of the novel prediction approach.

Furthermore, the results presented in Tables 2 for MDDR dataset indicated that this MaramalNet activity prediction technique showed the least variance for the Sensitivity, Specificity and the AUC values for all the activities classes in comparison to the classic NaiveB, RBFN and LSVM ML algorithms, indicating that the deep learning process should be considered as a novel, promising and interesting method for predicting the activities of chemical compounds.

5. CONCLUSION

In this study, the researchers investigated the deep convolutional networks (having up to 9 weight layers) for predicting the activities and for the ligand-based targets. They demonstrated that there was a lower representation depth for the prediction accuracy. They also proposed a novel mol2matrix technique, which was less overlapped and could segregate the biological activities of the molecules. Thereafter, they applied the new MaramalNet technique on the popular datasets and compared their performance with 3 standard ML algorithms. All experiments indicated that the MaramalNet algorithm exhibited interesting prediction rates (where the highly diverse dataset showed 88.01% accuracy, while a low diversity dataset showed 98% accuracy). Furthermore, the experiments also indicated that this novel MaramalNet algorithm showed an effective performance for the homogeneous datasetsbut showed a lower performance against the structurally heterogeneous datasets. Hence, the researchers have presented MaramalNet as a stable and convenient activity prediction approach for the unknown target chemical compounds. However, this area still needs to be explored further and better accuracy prediction techniques have to be developed for the highly diverse activity classes.

ACKNOWLEDGMENT

This work is supported by the Ministry of Higher Education (MOHE) and the Research Management Centre (RMC) at the Universiti Teknologi Malaysia (UTM) under the Research University Grant Category (VOT Q.J130000.2528.16H74 and R.J130000.7828.4F985). Also, we would like to thank Prof Sigeru Omatu for his great feedback during the early discussions about conducting this research.

REFERENCES

[1] A. Ammar, L. Valérie, J. Philippe, S. Naomie, and P. Maude, “Prediction of new bioactive molecules using a Bayesian belief network,” J. Chem. Inf. Model., vol. 54, no. 1, pp. 30–36, 2014.

[2] K. Barakat, “Computer-Aided Drug Design,” J. Pharm. Care Heal. Syst., vol. 1, no. 4, pp. 1–2, 2014.

[3] D. de la Iglesia, M. Garcia-Remesal, G. de la Calle, C. Kulikowski, F. Sanz, and V. Maojo, “The impact of computer science in molecular medicine: Enabling high-throughput research,” Curr. Top. Med. Chem., vol. 13, no. 5, pp. 526–575, 2013.

[4] S. Kothiwale, C. Borza, A. Pozzi, and J. Meiler, “Quantitative structure–activity relationship modeling of kinase selectivity profiles,” Molecules, vol. 22, no. 9, pp. 1–11, 2017.

[5] L. Wang et al., “Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field,” J. Am. Chem. Soc., vol. 137, no. 7, pp. 2695–2703, 2015.

[6] A. Vaidya, S. Jain, S. Jain, A. K. Jain, and R. K. Agrawal, "Quantitative Structure-Activity Relationships: A Novel Approach of Drug Design and Discovery," J. Pharm. Sci. Pharmacol., vol. 1, no. 3, pp. 219–232, 2014.

[7] C. H. Andrade, K. F. M. Pasqualoto, E. I. Ferreira, and A. J. Hopfinger, "4D-QSAR: Perspectives in drug design," Molecules, vol. 15, no. 5, pp. 3281–3294, 2010.

[8] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-based machine learning methods for predicting drug-target interactions: a brief review.," Brief. Bioinform., vol. 15, no. 5, p. bbt056-, 2013.

[9] F. Luan, T. Wang, L. Tang, S. Zhang, and M. Natália Dias Soeiro Cordeiro, "Estimation of the toxicity of different substituted aromatic compounds to the aquatic ciliate tetrahymena pyriformis by QSAR approach," Molecules, vol. 23, no. 5, 2018.

[10] C. F. Lagos, G. F. Segovia, N. Nu ez-Navarro, M. A. Faúndez, and F. C. Zacconi, "Novel FXa inhibitor identification through integration of ligand- and structure-based approaches," Molecules, vol. 22, no. 10, 2017.

[11] D. Filimonov, V. Poroikov, Y. Borodina, and T. Gloriozova, "Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors," J. Chem. Inf. Comput. Sci., vol. 39, no. 4, pp. 666–670, 1999.

[12] P. Willett, D. Wilton, B. Hartzoulakis, R. Tang, J. Ford, and D. Madge, "Prediction of ion channel activity using binary kernel discrimination," J. Chem. Inf. Model., vol. 47, no. 5, pp. 1961–1966, 2007.

[13] B. Chen, C. Mueller, and P. Willett, "Evaluation of a Bayesian inference network for ligand-based virtual screening," J. Cheminform., vol. 1, no. 1, pp. 1–10, 2009.

[14] X. Xia, E. G. Maliski, P. Gallant, and D. Rogers, "Classification of kinase inhibitors using a Bayesian model," J.Med.Chem., vol. 47, pp. 4463–4470, 2004.

[15] D. a Winkler and F. R. Burden, "Application of neural networks to large dataset QSAR, virtual screening, and library design.," Methods Mol. Biol., vol. 201, pp. 325–367, 2002.

[16] K. Kawai, S. Fujishima, and Y. Takahashi, "Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines," J. Chem. Inf. Model., vol. 48, no. 6, pp. 1152–1160, 2008.

[17] Y. LeCun, B. Yoshua, and H. Geoffrey, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[18] Y. Bengio, Learning Deep Architectures for AI, vol. 2, no. 1. 2009.

[19] A. Gupta, H. Wang, and M. Ganapathiraju, "Learning structure in gene expression data using deep architectures, with an application to gene clustering," 2015 IEEE Int. Conf. Bioinforma. Biomed., pp. 1328–1335, 2015.

[20] L. a Gatys, A. S. Ecker, and M. Bethge, "A Neural Algorithm of Artistic Style," arXiv Prepr., pp. 1–16, 2015.

[21] H. Wang, A. Meghawat, L.-P. Morency, and E. P. Xing, "Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis," vol. 1, 2016.

[22] H. Wang and B. Raj, "On the Origin of Deep Learning," Arxiv, pp. 1–72, 2017.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Adv. Neural Inf. Process. Syst., pp. 1–9, 2012.

[24] T. Sainath, A. R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp. 8614–8618, 2013.

[25] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," Proc. 27th Int. Conf. Mach. Learn., no. 3, pp. 807–814, 2010.

[26] Chen, Yu-Hsin and Krishna, Tushar and Emer, Joel and Sze, Vivienne, "Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks," in IEEE International Solid-State Circuits Conference, ISSCC 2016, Digest of Technical Papers, 2016, pp. 262–263.

[27] J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik, "Deep neural nets as a method for quantitative structure-activity relationships," J. Chem. Inf. Model., vol. 55, no. 2, pp. 263–274, 2015.

[28] G. E. Dahl, N. Jaitly, and R. Salakhutdinov, "Multi-task Neural Networks for QSAR Predictions," pp. 1–21, 2014.

[29] T. Unterthiner, A. Mayr, G. Klambauer, and S. Hochreiter, "Toxicity Prediction using Deep Learning," 2015.

[30] T. Unterthiner, A. Mayr, G. Klambauer, M. Steijaert, J. K. Wegner, and H. Ceulemans, "Deep Learning as an Opportunity in Virtual Screening," Deep Learn. Represent. Learn. Work. NIPS 2014, pp. 1–9, 2014.

[31] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively Multitask Networks for Drug Discovery," no. Icml, 2015.

[32] Y. Wang et al., "PubChem's BioAssay database," Nucleic Acids Res., vol. 40, no. D1, pp. D400--D412, 2011.

[33] A. P. Bento et al., "The ChEMBL bioactivity database: an update," Nucleic Acids Res., vol. 42, no. D1, pp. D1083--D1090, 2014.

[34] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," J. Chem. Inf. Model., vol. 50, no. 5, pp. 742–754, 2010.

[35] D. Dana et al., "Deep Learning in Drug Discovery and Medicine; Scratching the Surface," Molecules, vol. 23, pp. 1–15, 2018.

[36] J. MLA and G. M. Maggiora, "Concepts and Application of Molecular Similarity," Wiley Interdiscip. Rev. Mol. Sci., vol. 50, pp. 376–377, 1990.

[37] Y. C. Martin, J. L. Kofron, and L. M. Traphagen, "Do structurally similar molecules have similar biological activity?," J. Med. Chem., vol. 45, no. 19, pp. 4350–4358, 2002.

[38] "Sci Tegic Accelrys Inc." [Online]. Available: http://accelrys.com/products/collaborative-science/databases/bioactivity-databases/mddr.html.

[39] J. J. Sutherland, L. a. O'Brien, and D. F. Weaver, "Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure-Activity Relationships," J. Chem. Inf. Comput. Sci., vol. 43, no. 6, pp. 1906–1915, 2003.

[40] "Sutherland dataset." [Online]. Available: http://cdb.ics.uci.edu/cgibin/LearningDatasetsWeb.py.

[41] H. Hentabli, S. Naomie, and F. Saeed, "AN ACTIVITY PREDICTION MODEL USING SHAPE-BASED DESCRIPTOR METHOD," J. Teknol., vol. 1, pp. 1–8, 2016.

[42] H. Hentabli, N. Salim, A. Abdo, and F. Saeed, "LINGO-DOSM : LINGO for Descriptors of Outline," Intell. Inf. Database Syst. Springer Berlin Heidelb., pp. 315–324, 2013.

[43] H. Hentabli, N. Salim, A. Abdo, and F. Saeed, "LWDOSM : Language for Writing Descriptors," Adv. Mach. Learn. Technol. Appl. Springer Berlin Heidelb., pp. 247–256, 2012.

[44] H. Hentabli, F. Saeed, A. Abdo, and N. Salim, "A new graph-based molecular descriptor using the canonical representation of the molecule," Sci. World J., vol. 2014, 2014.

[45] T. Mikolove, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," pp. 1–9, 2013.

[46] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," pp. 1–12, 2013.

[47] R. Benigni, A. Giuliani, R. Franke, and A. Gruska, "Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines," Chem. Rev., vol. 100, no. 10, pp. 3697–3714, 2000.

[48] P. Ertl, "An algorithm to identify functional groups in organic molecules," J. Cheminform., vol. 9, no. 1, pp. 1–7, 2017.

[49] E. L. Schymanski et al., "Critical Assessment of Small Molecule Identification 2016: automated methods," J. Cheminform., vol. 9, no. 1, pp. 1–21, 2017.

[50] M. He, Q. Yang, A. Norvil, D. Sherris, and H. Gowher, "Characterization of Small Molecules Inhibiting the Pro-Angiogenic Activity of the Zinc Finger Transcription Factor Vezf1," Molecules, vol. 23, no. 7, p. 15, 2018.

[51] J. Kazius, R. McGuire, and R. Bursi, "Derivation and validation of toxicophores for mutagenicity prediction," J. Med. Chem, vol. 48, pp. 312–320, 2005.

[52] K. Hansen et al., "Benchmark data set for in silico prediction of Ames mutagenicity," J. Chem. Inf. Model., vol. 49, no. 9, pp. 2077–2081, 2009.

[53] F. Saeed and N. Salim, "Using soft consensus clustering for combining multiple clusterings of chemical structures," J. Teknol. (Sciences Eng., vol. 63, no. 1, pp. 9–11, 2013.

[54] V. GUPTA, "Image Classification using Convolutional Neural Networks in Keras," 2017. [Online]. Available: https://www.learnopencv.com/image-classification-using-convolutional-neural-networks-in-keras/.

[55] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[56] G. H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," pp. 338–345, 2013.

[57] C. CHIH-CHUNG, "LIBSVM : A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, p. 27:1-27:27, 2011.

[58] G. Bugmann, "Normalized Gaussian radial basis function networks," Neurocomputing, vol. 20, no. 1–3, pp. 97–110, 1998.

[59] F. Bastien et al., "Theano: new features and speed improvements," pp. 1–10, 2012.

[60] F. Chollet, "Keras Documentation," Keras.Io, 2015.

[61] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," 2012.