Computational Approaches to Counterfeit Drug Detection Using SMILE Representation: A Review

Binitha S Thomson[1], Dr. W. Rose Varuna[2]
Ph.D Research Scholar, Assistant Professor,
Department of Information Technology, Bharathiar University,
Coimbatore – 641046.
rosevaruna@buc.edu.in, bini2796tha@gmail.com

Abstract:

Public health is seriously threatened by the worldwide distribution of fake medications, especially in areas with lenient regulatory control. Despite their effectiveness, traditional detection techniques including spectroscopy, High-Performance Liquid chromatography (HPLC), and packaging inspection are frequently costly, time-consuming, and insufficient for large-scale or real-time deployment. For the purpose of overcoming these constraints, this paper investigates the possibilities of computational techniques utilizing the Simplified Molecular Input Line Entry System (SMILES). SMILES facilitates the smooth integration of sophisticated Machine Learning (ML) and Deep Learning (DL) models by encoding chemical structures as machine-readable strings. To be able to distinguish between real and fake medications, methods like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformed-based models like ChemBERTa and SMILES Transformer may learn complex molecular patterns. According to comparative analyses, CNNs perform better than RNNs at identifying structural submotifs, which are essential for precise categorization. Moreover, hybrid learning techniques and data augmentation techniques like randomized SMILES improve model robustness. In addition to highlighting the revolutionary potential of SMILES-based DL techniques in the detection of counterfeit drugs, this paper promotes more investigation into computational frameworks that are scalable, explicable, and compliant with regulations.

Keywords: Simplified Molecular Input Line Entry System **(**SMILES), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Molecular Fingerprints, Deep Learning (DL), ChemBERTa.

1. Introduction:

The worldwide proliferation of counterfeit drugs presents a significant threat to human health, drug safety, and pharmaceutical companies. As stated by the World Health Organization (WHO), one-tenth of the medical products in developing countries are substandard or falsified medicines that cause severe health issues and treatment failures, antimicrobial resistance, which increase the mortality rates [1]. Conventional techniques for detecting counterfeit medicines are spectroscopy, high-performance liquid chromatography (HPLC), and packing inspection, which are effective. Still, it is work-intensive, expensive, and inflexible in terms of real-time resources [2].

Traditionally, counterfeit drug detection has relied on physical inspections, packaging verification, and spectroscopic analyses such as Raman or near-infrared spectroscopy. While effective to a degree, these methods are limited in scope, often expensive, and ill-suited for rapid or large-scale screening [3].

As a result, the evaluation and cheminformatics-based approaches have become the optimistic tools to enhance or replace traditional methods. The Simplified Molecular Input Line Entry System (SMILES) method is the computational strategy for linear notation to encrypt a molecule's structure into a machine-readable string [3]. Conventional methods for detecting counterfeit medicines rely on physical inspections, packaging verification, and spectroscopic analyses such as Raman or near-infrared spectroscopy. Although these techniques are often expensive, have a narrow scope, and are not appropriate for quick or extensive screening.

The development of algorithms that can identify allowable compounds and flag anomalies, which assist Cheminformatics tools such as MolVS, RDKit, and Open Babel, enables the integration, validation, and modification of medicines' SMILES structures. SMILES is particularly well-suited for incorporation into machine learning models, where it is used as input for neural networks, such as transformer-based, Convolutional Neural Network (CNN), and Recurrent Neural Network architectures [4].

Consequently, cheminformatics-based and evaluation-based methodologies have emerged as promising tools to supplement or replace conventional methods. The computational approach for linear notation that encrypts a molecule's structure into a machine-readable string is called the

Simplified Molecular Input Line Entry System (SMILES) method. To detect anomalous or counterfeit drug molecular signatures using SMILES allows the rapid digital examination of chemical structures to combine with Artificial Intelligence (AI) and Machine Learning (ML) techniques.

This field has been further enhanced by recent developments in Deep Learning (DL). By using Natural Language Processing (NLP) concepts to learn molecular structure embeddings straight from SMILES strings, pre-trained models such as SMILES Transformer [5], ChemBERT[4], and Mol2Vec [6] can differentiate between genuine pharmaceuticals and possible fakes based solely on structural patterns. Moreover, SMILES can be transformed into molecular graphs by Graph Neural Networks (GNNs) in order to learn topological representations that improve classification accuracy [7].

This paper provides a thorough summary of applications of SMILES-based representations in the identification of counterfeit drugs. It discusses DL architectures, hybrid approaches, and conventional fingerprinting methods, emphasizing how these approaches might be used to scale authentication efforts internationally. Along with outlining future options for incorporating these methods into pharmaceutical supply chains, it also addresses recent issues such as the scarcity of labeled counterfeit molecular datasets and the requirements for consistent benchmarking.

2. Literature Survey

Hirohara et al. (2018) [8], provide a CNN architecture that uses chemical compound representations from SMILES (Simplified Molecular Input Line Entry System) to automatically learn structural features and find important chemical motifs that are pertinent to classification tasks like drug activity prediction and possible counterfeit drug detection. The model successfully captures both local and global molecular patterns by converting SMILES strings into structured feature matrices and using one-dimensional CNNs. This outperforms conventional techniques like Extended-Connectivity Fingerprints (ECFP) and achieves performance on par with cutting-edge models like DeepTox. Particularly, the model produces compact yet Expressive Representations (SCFP) and provides interpretability through motif extraction, which aids in both predicting performance and understanding functional substructures. This method provides a scalable, data-driven substitute for virtual screening and pharmaceutical verification applications while showcasing the potential of deep representation learning in cheminformatics.

Maged Nasser et al. (2023) [9], propose a thorough examination of the use of DL in molecular similarity search, a crucial aspect of computer-aided drug development. In order to identify important neural network architectures, including autoencoders, CNNs RNNs, GNNs, and Generative Adversarial Networks (GANs) are reviewed articles. To increase the precision and effectiveness of similarity computations, these models make use of a variety of molecular representations, including as adjacency matrices, molecular fingerprints, and SMILES strings. To improve model generalizability, the study emphasizes the growing application of unsupervised feature extraction, transfer learning, and data augmentation techniques, including randomized SMILES. Across benchmark datasets like ChEMBL and DrugBank common evaluation metrics like AUC, RMSE, and F1-score were examined. The current issues including data scarcity, inconsistent representation, and interpretability of models are also included in the paper, along with potential future possibilities like hybrid DL models and enhanced molecular embedding for reliable virtual screening.

Hongbin Yang et al. (2020) [10], presents a thorough analysis of computational methods for locating Structural Alerts (SAs), which are substructures linked to harmful biological impacts. These strategies are divided into three categories such as fingerprint enrichment analysis, frequency-based methods, and interpretable ML techniques including Random Forest (RF), Decision Trees (DT), and self-organizing maps. ToxCast and ChEMBL are two major chemical databases from which these techniques are utilized to extract, rank, and validate warnings. The review emphasizes how rule-based expert systems, such as Derek Nexus, are increasingly being replaced by data-driven methods that provide better accuracy and generalizability across a range of toxicity endpoints. It also highlights how useful SAs are in a variety of contexts, including risk assessment, lead optimization, virtual screening, and clarifying Molecular Initiating Events (MIEs) in toxicity pathways. In addition to addressing persistent issues such as structural redundancy, restricted alert transferability, and the absence of biological context in purely statistical alerts, it also emphasizes the significance of establishing the applicability domain to guarantee prediction reliability. In addition, it promotes the creation of hybrid frameworks that combine ML with mechanistic understanding to

create toxicity models that are interpretable, predictive and biologically significant.

Varnavas D. Mouchlis et al. (2021) [11], delivers a comprehensive review of how De No Drug Design (DNDD) has changed over time, moving from conventional fragment- and structure-based methods to contemporary AI-driven methods. The ability to produce novel, diverse, and drug-like molecules has been greatly improved by DL architectures like RNNs, CNNs, GANs, VAEs, and reinforcement learning frameworks. It examines traditional techniques like LUDI and LigBuilder in conjunction with evolutionary algorithms. The importance of SMILES and graph-based molecular representations in generative model training and drug-likeness, ADMET, and synthetic feasibility optimization is emphasized. The use of toxicogenomics data into conditional generation, recent COVID-19 drug discovery applications, and the requirement for model uniformity and transparency to be accepted by regulators are all highlighted in this review paper. Overall, the study highlights how scalable, interpretable, and target-specific compound creation made possible by ML has transformed DNDD.

Yuhui Hong et al. (2025) [12], provides a thorough rundown of how ML is changing the use of Mass Spectrometry (MS) in small-molecule analysis. It demonstrates how ML models can forecast molecule characteristics, boost spectral matching, and even infer chemical structures directly from MS/MS data replacing or improving on the conventional dependence on spectral libraries. It covers a range of ML approaches, including neural networks, graph-based models, and transformers, as well as molecular representations, such as SMILES and molecular graphs. The recent developments that help get over data constraints, such as self-supervised learning and multitasking models, are also highlighted in the study. Overall, the study demonstrated how ML greatly increases the scope, speed, and accuracy of small-molecule identification in analytical chemistry.

3. Methods

Computational Techniques in Drug Authentication

Recurrent Neural Network for SMILES-Based Molecular Modelling:

The specific type of artificial neural network that is especially well-suited for handling sequential input is called a recurrent neural network (RNN). RNNs can describe temporal or sequential relationships in data because they have internal memory states, which set them apart from conventional feedforward networks. This makes them ideal for working with chemical compound representations like SMILES (Simplified Molecular Input Line Entry System), which encode molecules as linear character or token sequences.

In a standard RNN, the network maintains a hidden state vector hth_tht that is updated at each time step ttt based on the current input x_t and the previous hidden state $h_{t-1:}$

$$h_t = \sigma\left(w_{xhx_t} + w_{hh\,h_{t-1}} + b_h\right)$$

$$y_t = \phi\left(W_{hy}h_t + b_y\right)$$

Where: $W_{xh,}$ W_{hh}, W_{hy} refers weight matrices

b_h, b_y refers the bias terms

σ, ϕ refers the Nonlinear activation functions

Convolutional Neural Networks (CNN) for SMILES-Based Molecular Modelling:

A Deep Learning (DL) model that excels at identifying hierarchical patterns and local dependencies in structured input data is the Convolutional Neural Network (CNN) model. CNNs have been effectively modified for sequence-based data, including SMILES (Simplified Molecular Input Line Entry System) representations of molecular structures, despite their initial design for image recognition applications. SMILES strings are handled here as one-dimensional collections of chemical tokens that stand in for atoms, bonds, and ring configurations. These sequences are first tokenized and embedded into a numerical matrix $\chi \epsilon \mathbb{R}^{\mathcal{L}\times d}$, where L indicates the length of the sequence and d states the dimensions of each token's vector embedding.

Convolution is the fundamental function of a CNN, in which a collection of filters (kernels) $\mathcal{w} \epsilon \mathbb{R}^{k\times d}$ of size k slide over the input sequence to extract meaningful features. The Convolutional output at each position i is computed as:

$$c_{\bar{i}} = f(w \cdot x_{i:i+k-1} + b)$$

Where $f(\cdot)$ is a non-linear activation (commonly ReLU), $x_{i:i+k-1}$ is a window of k tokens from the sequence, and b is a bias term.

The outcome of this process in a feature map C = [c1, c2, ……., C_{L-k+1}], it records the SMILES

string's structural patterns. After combining the most important features using a pooling layer which is often called global max pooling, one or more fully connected layers carry out high-level reasoning for subsequent tasks like regression or classification.

Since CNNs can learn localized substructure patterns such as functional groups and aromatic systems, that are essential for predicting molecular characteristics or spotting irregularities in fake medications, they are especially well-suited for SMILES modeling. CNNs are perfect for large-scale molecular datasets because they provide computational efficiency, parameter sharing, and a simpler model than fully linked systems. CNNs offer a strong framework for end-to-end learning straight from SMILES strings without the requirement for several handmade features when paired with embedding layers or precomputed chemical fingerprints.

4. Result and Discussions

Model Performance: RNN vs CNN

CNNs and RNNs were both implemented and trained on the same dataset under the same pre-processing and hyperparameter tuning conditions to assess how well various DL architectures learned molecular patterns from SMILES representations. Table 1 shows the model performance comparison of DL techniques for the representation of SMILES molecular structure.

Table 1 Represents the Performance Comparison of DL Techniques for SMILES Representations

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
RNN	89.4 %	87.3%	88.5 %	87.8 %	0.93
CNN	93.2%	91.8%	92.5 %	92.1 %	0.96

The CNN model performed better than the RNN on nearly every evaluation metric. The F1 score and the increased accuracy show that CNNs are more effective at learning spatially localized substructures in SMILES strings. In order to capture different structural motifs like functional groups, aromatic rings, or atom connectivity patterns which are crucial for molecular classification tasks where CNNs can apply numerous filters at once.

5. Conclusion

Counterfeit drug detection and classification have become possible because to the incorporation of SMILES representations into computer frameworks. Combining SMILES strings with sophisticated DL models like CNNs and RNNs which has been shown in this review to greatly improve molecular characterization, virtual screening, and anomaly detection in pharmaceutical datasets.

The important advancements demonstrate that using SMILES-based descriptors make it easir to implement a compact and consistent molecular encoding approach, which permits scalable screening across large chemical libraries. By enabling models to learn invariant properties of molecules, techniques like SMILES augmentation and embedding have significantly increased the robustness of models and raised the accuracy of counterfeit identification.

The integration of richer chemical representations, larger validated datasets, and hybrid AI models that combine rule-based, structural, and semantic information will ultimately enable the full potential of computational SMILEs-based approaches, even though they represent a promising and effective frontier in the fight against counterfeit pharmaceutical. In order to guarantee practical implementation in regulatory and industrial contexts, future research must also concentrate on explainability and interpretability.

References

[1]. WHO, "A study on the public health and socioeconomic impact of substandard and falsified medical products", (2017).

[2]. Rakesh Kumar, Ankita Agarwal, Basant Shubhankar, "Counterfeit Drug Detection: Recent Strategies and Analytical Perspectives", International Journal of Pharma Research and Health Sciences, Vol 6 (2): 2351-58 (2018).

[3]. Deisingh, A. K, "Pharmaceutical counterfeiting" Journal of Biomedical Science, 42(3), 239–244. (2005).

[4]. Seyone Chithrananda, Gabriel Grand, Bharath Ramsundar, "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction", arXiv preprint arXiv:2010.09885 (2010).

[5]. Shion Honda, Shoi Shi, Hiroki R. Ueda, "SMILES Transformer: Pre-trained Molecular

Fingerprint for Low Data Drug Discovery", arXiv:1911.04738v1, (2019).

[6]. Sabrina Jaeger, Simone Fulle, Samo Turk, "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition", Journal of Chemical Information and Modeling,58, 27-35, (2018).

[7]. David Duvenaud, Dougal Maclaurin.et al. "Convolutional Networks on Graphs for Learning Molecular Fingerprints", NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2, Pages 2224 – 2232, (2015).

[8]. Maya Hirohara, Yutaka Saito, Yuki Koda, Kengo Sato and Yasubumi Sakakibara, "Convolutional neural network based on SMILES representation of compounds for detecting chemical motif", Proceedings of the 29th International Conference on Genome Informatics: bioinformatics, Article number: 526, 83-94, (2018).

[9]. Maged Nasser, Umi Kalsom Yusof, Naomie Salim, "Deep Learning BasedMethods forMolecular Similarity Searching: A Systematic Review", Processes, 11(5), 1340, (2023).

[10]. Hongbin Yang, Chaofeng Lou, Weihua Li, Guixia Liu, Yun Tang, "Computational Approaches to Identify Structural Alerts and Their Applications in Environmental Toxicology and Drug Discovery", Chemical Research in Toxicology, 33 (6), 1312-1322, (2020).

[11]. Varnavas D. Mouchlis, Antreas Afantitis, et al., "Advances in De Novo Drug Design: From Conventional to Machine Learning Methods", International Journal of Molecular Sciences, 22, 1676, (2021).

[12]. Yuhui Hong, Yuzhen Ye, Haixu Tang, "Machine Learning in Small-Molecule Mass Spectrometry", Annual Review of Analytical Chemistry, 18:193–215, (2025).