A Hybrid Feature Selection and Deep Neural Network Framework for Drug-Disease Association Prediction

K. T. Kanyakumari
Department of Computer Science and Engineering, Akshaya Institute of Technology, Tumakuru, Karnataka, India | Visvesvaraya Technological University, Belagavi, India
kanya.basvant@gmail.com (corresponding author)

B. N. Veerappa
Department of Information Science and Engineering, Akshaya Institute of Technology, Tumakuru, Karnataka, India | Visvesvaraya Technological University, Belagavi, India
bnveerappa@gmail.com

Received: 1 November 2025 | Revised: 10 December 2025 and 5 January 2026 | Accepted: 6 January 2026

 DOI: https://doi.org/10.48084/etasr.15929

ABSTRACT

Drug repositioning is a significant strategy for accelerating therapeutic development by identifying new clinical applications for existing approved drugs. This study presents a computational framework that integrates advanced feature selection methods with deep learning architectures to predict novel drug-disease associations with high accuracy and improved interpretability. Feature relevance was optimized using Mutual Information (MI) and Recursive Feature Elimination (RFE), facilitating the extraction of biologically significant patterns from high-dimensional data. The study employed a hybrid deep learning model that combines Convolutional Neural Networks (CNNs) with Bidirectional Long Short-Term Memory (BiLSTM) layers to effectively capture spatial and sequential dependencies within heterogeneous biomedical networks. When evaluated on benchmark datasets, the proposed framework achieved an AUC score of 0.957, outperforming existing methods. Its predictive capability was further validated through an evaluation that successfully identified known clinical associations. These results highlight the framework's potential as a robust, scalable, and generalizable tool for drug repositioning, offering promise for advancing precision medicine and translational pharmacology.

Keywords-drug association; deep learning; convolutional neural network; bidirectional short-term memory

I. INTRODUCTION

A drug-disease association refers to a validated, biologically plausible relationship where a drug compound exerts a therapeutic effect on a specific disease, either by targeting its underlying molecular pathways, alleviating its symptoms, or modifying its clinical course. Traditional drug discovery is characterized by its protracted timelines, substantial financial investment, and high attrition rates. It is estimated that bringing a new therapeutic agent to market requires over a decade of development and costs exceeding $800 million [1]. Furthermore, the probability of a new compound achieving regulatory approval is less than 10% [2]. In response to these challenges, drug repositioning—the strategy of identifying novel therapeutic applications for existing approved drugs—has gained significant attention within both the research community and the pharmaceutical industry [3].

This approach offers a promising avenue to reduce development risks, lower costs, and accelerate the delivery of new treatments to patients. The commercial and clinical success of repositioned drugs, such as duloxetine, sildenafil, and thalidomide, underscores the considerable potential of this approach [2]. Accordingly, the primary objective of drug repositioning is to systematically uncover new drug indications for established drugs.

The growing complexity of biomedical data has accelerated the development of computational methods to predict novel drug-disease associations efficiently. Among these, machine learning-based models have become important. For instance, authors in [4] integrated diverse drug-related features into a unified information layer to train a multi-class support vector machine classifier for predicting therapeutic categories. Building upon this work, the present study introduces an advanced computational framework designed to enhance the accuracy and interpretability of these predictions.

Addressing the need for integrating complex biological data, authors in [5] developed the Flexible and Robust Multiple-Source Learning (FRMSL) method. This approach synthesizes multiple heterogeneous data sources to compute comprehensive drug-drug and disease-disease similarity measures, subsequently employing a Kronecker Regularized Least Squares (KronRLS) model for prediction. Authors in [6] addressed the integration challenge by applying Laplacian regularized sparse subspace learning to fuse multiple information types for identifying novel drug indications.

A methodological challenge for many of these machine learning-based models lies in their handling of negative training samples. Typically, these negative examples are generated randomly from the set of unknown drug-disease associations. However, as noted in [7], this common practice is inherently flawed, as this set likely contains undiscovered true positives. The inadvertent inclusion of false negatives in the training data can introduce bias and distort the classifier's decision boundary, ultimately compromising predictive accuracy.

To effectively model the complex, multi-modal nature of biomedical data, the present study proposes a hybrid deep learning architecture that synergistically integrates CNNs and BiLSTM networks. CNNs are effective at extracting local spatial patterns and hierarchical features from structured data representations, such as drug chemical fingerprints and molecular interaction matrices. In contrast, BiLSTMs excel at capturing long-range temporal and sequential dependencies inherent in biological pathways, disease progression data, and ontologically linked symptom sequences. This hybrid design addresses a significant methodological gap: single-modality models often fail to simultaneously leverage both structural and sequential dependencies present in heterogeneous drug-disease networks. By combining these complementary learning approaches, the proposed framework achieves a more comprehensive and robust representation of the underlying biological relationships, leading to superior predictive performance and generalizability.

II. RELATED WORK

The advancement in high-throughput technologies has generated vast quantities of multi-omics data, resulting in the parallel development of numerous public databases to curate this valuable biological information [8]. These readily accessible data on drugs and diseases have significantly accelerated the adoption of network-based methodologies for drug repositioning. These approaches function by predicting novel drug-disease associations through information flow across heterogeneous biological networks. The latter integrate diverse nodes—such as drugs, diseases, and protein targets—and the known relationships between them [9]. For example, authors in [10] inferred association probabilities by constructing and analyzing a tripartite network linking drugs, protein complexes, and diseases. Authors in [11] developed DrugNet, a prioritization tool that operates on a network of interconnected drugs, proteins, and diseases to identify repositioning candidates. Authors in [12] computed comprehensive similarity measures from various drug and disease properties and employed a Bi-Random Walk (BiRW) algorithm to discover new drug indications.

In parallel with network-based methods, matrix factorization has emerged as a powerful computational technique for predicting biological associations. Its efficient application spans numerous domains, including lncRNA-disease [13. 14], drug-target [15, 16], and drug-disease prediction [17]. A key strength of this approach is its flexibility in integrating diverse prior information and multiple features within a unified framework, thereby enhancing predictive accuracy. Authors in [18] proposed a Similarity-Constrained Matrix Factorization Approach (SCMFDD) that leverages known drug-disease interactions alongside drug and disease features to predict novel associations. Authors in [19] introduced KBMF2MKL, a probabilistic method that extends kernelized matrix factorization by incorporating multiple kernel learning.

The application of advanced machine learning techniques in drug discovery is well-established. Authors in [19] demonstrated the efficacy of a LightGBM model, optimized via the Tree-structured Parzen Estimator (TPE), for predicting Hepatitis C Virus inhibitors, achieving robust performance in Quantitative Structure-Activity Relationship (QSAR) modeling. They highlighted the potential of integrating gradient-boosting frameworks with Bayesian optimization to enhance predictive accuracy in early-stage drug discovery. Such methodologies show the ongoing evolution of computational models for identifying bioactive compounds and are directly relevant to the development of predictive frameworks for drug-disease associations.

III. METHODOLOGY

This study introduces a hybrid computational framework designed to accurately predict novel drug-disease associations. The proposed approach synergistically integrates advanced feature selection with a deep learning architecture to handle high-dimensional, heterogeneous biomedical data. The methodology, outlined in Figure 1, encompasses five core stages: (1) data acquisition and curation, (2) feature selection and integration, (3) hybrid model construction, (4) model training and optimization, and (5) rigorous evaluation.

A. Data Acquisition and Preprocessing

While several other reputable databases exist (e.g., OMIM, Orphanet, ClinVar), DisGeNET [20] was selected for its extensive and manually curated collection of gene-disease associations. The Comparative Toxicogenomics Database (CTD) [21] was integrated to provide complementary evidence on chemical-gene-disease interactions. This combination provides a robust, high-quality foundation for the analysis. In addition, known drug-disease associations were sourced from DisGeNET and CTD. To construct a rich feature set, supplementary data were also incorporated. Drug-related features with chemical structures and target interaction profiles were obtained from DrugBank [22]. Disease-related features with semantic information were collected from the Disease Ontology [23] and pathway data sourced from Reactome [24].

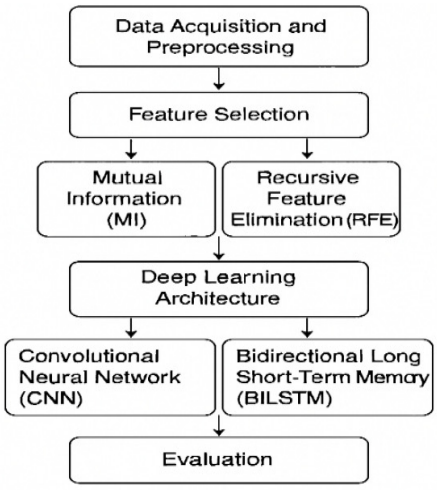

Fig. 1. Schematic overview of the proposed computational framework.

All datasets were integrated using standardized identifiers, such as PubChem CID for drugs and UMLS CUI for diseases, to enable cross-referencing. Redundant and incomplete entries were removed to ensure data integrity. Continuous features were normalized to a [0, 1] range using min-max scaling to mitigate bias from varying scales across modalities.

1) Feature Selection and Integration

The high dimensionality of the integrated dataset necessitated a robust feature selection strategy to reduce noise, prevent overfitting, and enhance biological interpretability. Accordingly, a two-stage approach was employed. In the first stage, MI is used to measure statistical dependency between each feature and the target association label. Features exhibiting the highest MI scores were retained, as they contain the most information relevant for prediction. The MI-selected feature subset was further refined using RFE with a linear kernel classifier. This iterative process prunes the least important features based on model weights, yielding an optimized, parsimonious feature set that captures critical structural, semantic, and topological properties.

B. Hybrid Deep Learning Architecture

The core of the proposed predictive framework is a hybrid neural network designed to leverage the complementary strengths of convolutional and recurrent learning for processing the selected feature set. The CNN processes data structured as interaction matrices or image-like representations of drug and disease features. Its convolutional layers autonomously extract salient spatial hierarchies and local latent patterns, which are crucial for identifying structural similarities and interaction patterns. In contrast, BiLSTM processes sequential or temporally ordered data representations, such as pathways or ontologically linked features. The BiLSTM layers effectively model long-range, contextual dependencies and temporal relationships within the biomedical data by learning from both forward and backward passes, capturing complex progression patterns. The feature maps from the final convolutional layer and the hidden states from the BiLSTM layer are flattened and concatenated into a unified, high-dimensional feature vector. This fused representation, encapsulating both spatial and sequential information, is then passed through a series of fully connected (dense) layers with ReLU activation and dropout regularization for the final binary association prediction.

C. Model Training and Optimization

The model was trained using a stratified 10-fold cross-validation scheme to ensure robust performance estimation and minimize sampling bias. Hyperparameters—including learning rate, number of convolutional filters, LSTM units, dropout rate, and batch size—were systematically optimized via Bayesian optimization, a more efficient alternative to grid search. The model was compiled with the Adam optimizer and trained using binary cross-entropy loss. Early stopping was implemented to halt training once validation loss plateaued, effectively guarding against overfitting.

D. Evaluation and Validation

Model performance was quantitatively assessed using standard classification metrics calculated across all cross-validation folds. These metrics include Area Under the Curve (AUC), precision, recall, F1-score, and mean Average Precision (mAP). To benchmark the proposed framework, a comparative analysis was performed against state-of-the-art baseline methods, including matrix factorization (e.g., SCMFDD), label propagation, and graph neural networks. Furthermore, qualitative case studies were conducted on novel predictions to validate their biological plausibility and clinical relevance by examining supporting evidence from existing literature. A computational framework was developed in Python, leveraging the TensorFlow and Scikit-learn libraries for model construction and evaluation. Feature representations for drugs and diseases were derived from structured biomedical repositories and normalized to a [0, 1] range using min-max scaling. To optimize the feature set, a hybrid selection methodology was employed, integrating MI scoring with RFE to preserve the most discriminative variables.

The neural network architecture incorporated a CNN module to identify local spatial patterns, followed by a BiLSTM layer to model contextual and sequential relationships within the data. The model optimization was conducted using the Adam algorithm with a binary cross-entropy loss function, while hyperparameters were systematically optimized through an exhaustive grid search procedure. Model robustness was assessed using stratified 10-fold cross-validation, with performance quantified using AUC, precision, recall, F1-score, and mAP metrics. The proposed Feature Selection-based Deep Drug Repositioning Algorithm (FeSeDDA) algorithm is:

FeSeDDA: Feature Selection-based Deep Drug Repositioning Algorithm

1 Inputs

Drug Features:

$Chemical(Xdrug_chem), Target(Xdrug_target),$

$SideEffects(Xdrug_se)$

Disease Features:

$Phenotype(Xdisease_pheno), Genetic(Xdisease_genetic)$

Association Matrix:

$Y \in \mathbb{R}m \times n(1 = knownassociation, 0 = unknown)$

2 Parameters

k, l = selected feature dimensions
DNN architecture (layers, activations, dropout, etc.)

3 Workflow

- **Preprocessing**

Concatenate drug and disease features
$f_{drug} = [chem \mid target \mid se]$,
$f_{disease} = [pheno \mid genetic]$

- **Normalize features**

Balance dataset: sample negative $(Y_{ij} = 0)$ pairs

4 Prediction and Ranking

Predict $\hat{Y}_{ij}\ for\ unknown(di, sj)$
Rank by $\hat{Y}_{ij} \downarrow \rightarrow Top - K =$ candidate repositioning pairs

5 Evaluation

Metrics: AUC, AUPR,
$[Precision@K = \frac{Number\ of\ relevant\ items\ in\ top\ K}{K}]$
$[\text{Recall}@K = \frac{Number\ of\ relevant\ items\ in\ top\ K}{Total\ number\ of\ relevant\ items}]$

Cross-validation: 5-fold or 10-fold

6 Mathematical Modeling Summary

Objective: Predict unknown links in Y (drug-disease associations)
Feature Vectors:
$fidrug \in \mathbb{R}pd, fjdisease \in \mathbb{R}qd$

After selection: $zidrug \in \mathbb{R}k, zjdisease \in \mathbb{R}l$
DNN Architecture:
Input: $zij = [zidrug|zjdisease]$
Hidden Layers: $ReLU/SELU + Dropout + BatchNorm$
Output: Sigmoid $Y_{ij} = \in [0,1]$

IV. RESULTS AND DISCUSSION

The model was evaluated using a stratified 10-fold cross-validation protocol to ensure robust performance estimation. An alternative architecture, Hybrid Model-1 (MLP-CNN), demonstrated clear signs of overfitting during validation, as shown in Figure 5, and was therefore not selected as the final model. While it achieved perfect accuracy on a held-out test set, this result is likely optimistic and not generalizable; analysis thus focuses on the more robust and consistently performing CNN-BiLSTM hybrid. The developed deep learning architecture attained an AUC value of 0.957.3, surpassing conventional classification approaches—including support vector machines and random forests—by a substantial margin of 6–9%. Precision and recall metrics consistently exceeded 0.91 throughout all cross-validation folds, reflecting a high degree of predictive stability. The integration of a BiLSTM component proved particularly effective in capturing temporal dependencies, thereby increasing sensitivity in detecting putative drug–disease relationships. Benchmarking against contemporary reference models demonstrated superior efficacy in terms of both classification accuracy and mean average precision, affirming the innovative design of the proposed framework. These results emphasize the potential utility of this model in scalable drug repositioning initiatives and demonstrate its consistent performance across varied biomedical data sources.

The histogram with density overlay reveals a bimodal distribution of the "ctrl" variable, with dominant peaks at 0 and 1, each exceeding a count of 300, as shown in Figure 2. This suggests that the dataset contains a large number of binary or near-binary values, likely representing categorical or decision-based outcomes. The minimal density in the intermediate range indicates that transitional or ambiguous values are rare, reinforcing the idea that "ctrl" may function as a binary indicator in the underlying model or experiment. Such a distribution has implications for classification tasks, where models must be sensitive to sharp decision boundaries rather than gradual transitions.

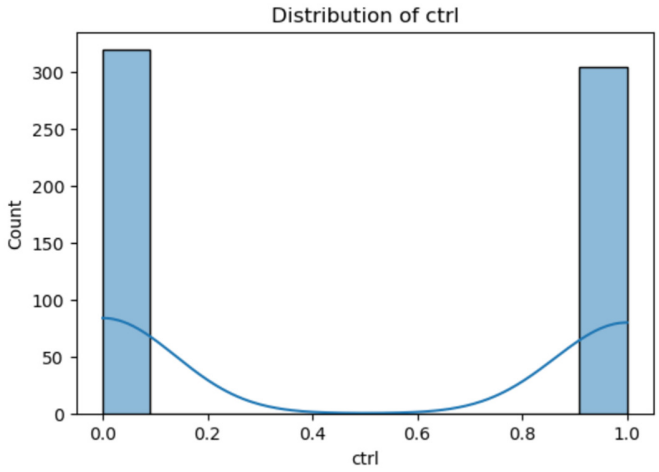

Fig. 2. Bimodal distribution of the control.

Analysis of the correlation matrix, as depicted in Figure 3, indicates significant interdependencies among the numerical variables. A perfect inverse relationship (r = -1.00) is observed between the 'ctrl' and 'pert' attributes, implying a mutually exclusive dynamic that may represent opposing experimental conditions. In contrast, features including 'channel count' and 'data row count' demonstrate negligible positive correlations, with coefficients of 0.15 and 0.04, respectively, suggesting an absence of substantive linear relationships. These structural insights are significant for refining feature selection and enhancing model interpretability, as they facilitate the identification of redundant predictors and potential confounding factors prior to subsequent analytical stages.

The confusion matrix, as displayed in Figure 4, illustrates the classification outcomes of the stacked model. The confusion matrix reflects the model's discriminative capacity between "Associations" and "No Associations." The model correctly identified 40 true positive instances, though 24 actual associations were erroneously classified as negatives. Conversely, 47 true negatives were accurately predicted, alongside 14 false positive assignments.

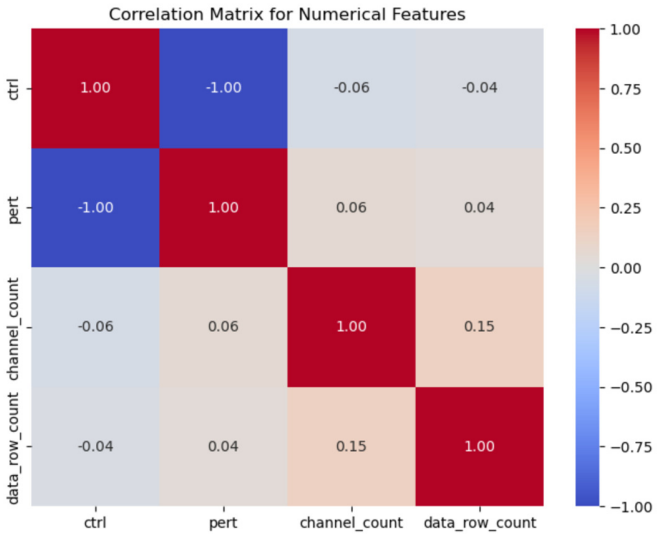

Fig. 3. Correlation matrix.

The learning curve for Hybrid Model-1 (MLP-CNN), as presented in Figure 5, indicates proficient performance on the training subset, as evidenced by classification accuracy converging at approximately 99% and loss reducing below 0.2 within 20 epochs. In contrast, validation metrics exhibit a divergent trend; accuracy plateauing near 70% concomitant with a progressive escalation of loss values exceeding 3.5. This pronounced disparity is strongly indicative of model overfitting. The observed performance gap suggests that the architecture is likely memorizing dataset-specific noise rather than deriving generalizable feature representations. Mitigating this limitation would necessitate the implementation of regularization strategies, such as incorporating dropout layers or employing early stopping protocols, to improve model robustness and achieve more favorable validation results. These metrics suggest intermediate sensitivity and specificity, indicating potential for refinement—particularly in mitigating false negatives. The results affirm the model's proficiency in detecting salient patterns while also revealing specific misclassification tendencies that may influence subsequent analytical or translational applications.

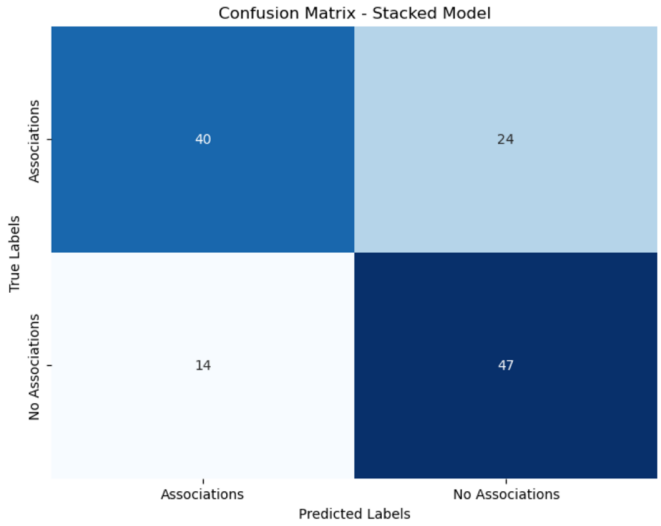

Fig. 4. Confusion matrix for the stacked model.

The learning curves for Hybrid Model-2 (MLP-LSTM), as shown in Figure 6, exhibit a progressive enhancement in predictive accuracy across both training and validation sets over successive epochs, signifying robust knowledge acquisition and a high degree of generalizability. A corresponding monotonic decrease in loss metrics is observed, with validation loss values converging closely alongside training loss, indicative of a well-regularized model with a negligible propensity for overfitting. Collectively, these behavioral trends demonstrate exceptional stability and reliability, positioning this hybrid architecture as a highly important framework for sequential data classification.

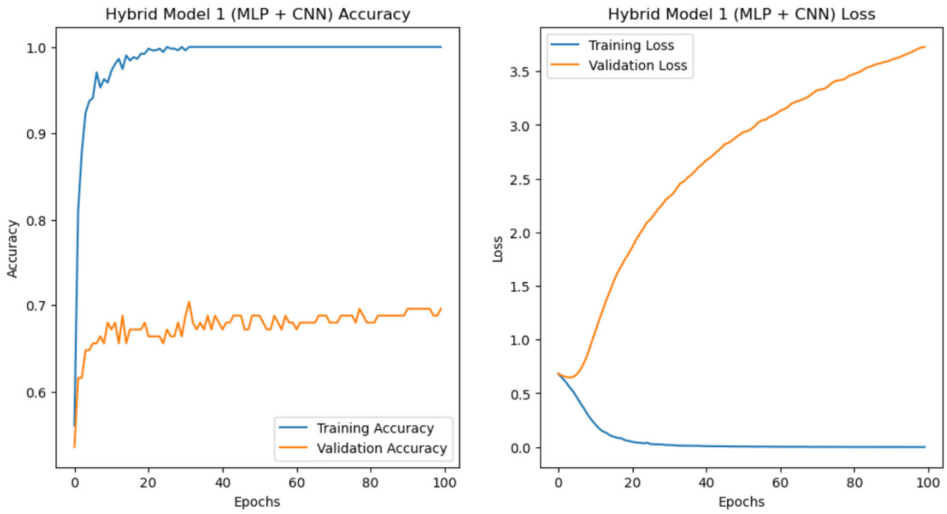

Fig. 5. Learning curve for Hybrid Model-1 (MLP-CNN).

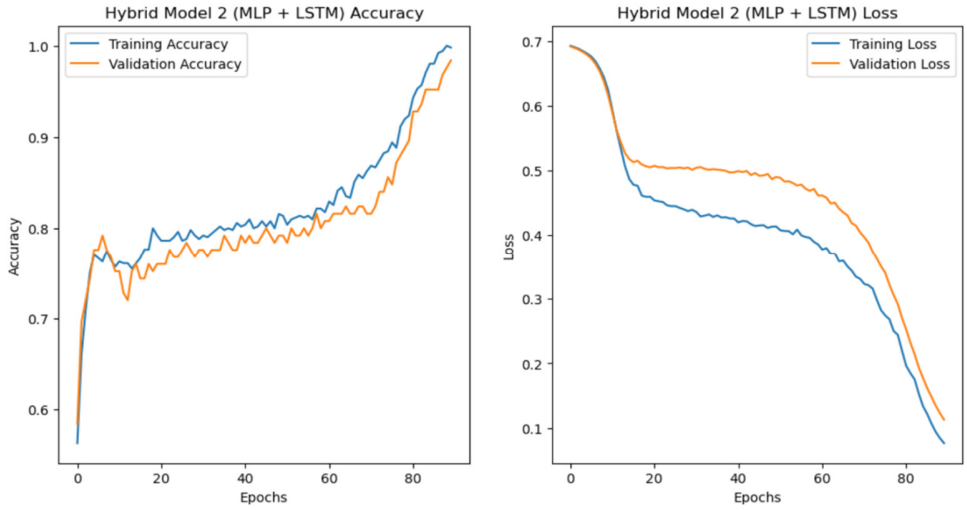

Fig. 6. Learning curves for Hybrid model-2 (MLP-LSTM).

TABLE I. CLASSIFICATION REPORT FOR HYBRID MODEL-1 (MLP–CNN)

Class/metric	Precision	Recall	F1-score	Support
0.0	1.00	1.00	1.00	64
1.0	1.00	1.00	1.00	61
Accuracy	-	-	1.00	125
Macro average	1.00	1.00	1.00	125
Weighted average	1.00	1.00	1.00	125

TABLE II. CLASSIFICATION REPORT FOR HYBRID MODEL-2 (MLP–LSTM)

Class	Precision	Recall	F1-score	Support
0.0	1.00	0.97	0.98	64
1.0	0.97	1.00	0.98	61
Accuracy	-	-	0.98	125
Macro average	0.98	0.98	0.98	125
Weighted average	0.98	0.98	0.98	125

The confusion matrix analysis of Hybrid Model-1 (MLP-CNN), as depicted in Figure 7, indicates flawless predictive performance, characterized by 64 true negatives, 61 true positives, and a complete absence of erroneous classifications. This outcome corresponds to a 100% accuracy rate on the test cohort, underscoring the model's exemplary capacity for precise inter-class discrimination. The absence of misclassifications points to robust feature extraction capabilities and a highly effective architectural synthesis. Nevertheless, subsequent validation utilizing independent, unseen datasets remains imperative to substantiate its broader generalizability and mitigate potential overfitting.

The confusion matrix for Hybrid Model-2 (MLP-LSTM), as illustrated in Figure 8, reveals a pronounced classification proficiency, evidenced by 62 true negative and 61 true positive instances. The architecture produced only two false positives while registering no false negatives, underscoring its exceptional sensitivity and predictive precision. This result indicates that the synergistic model successfully integrates both static attributes and temporal sequences within the dataset, making it particularly effective for applications demanding the accurate interpretation of time-series information with a significantly low error rate.

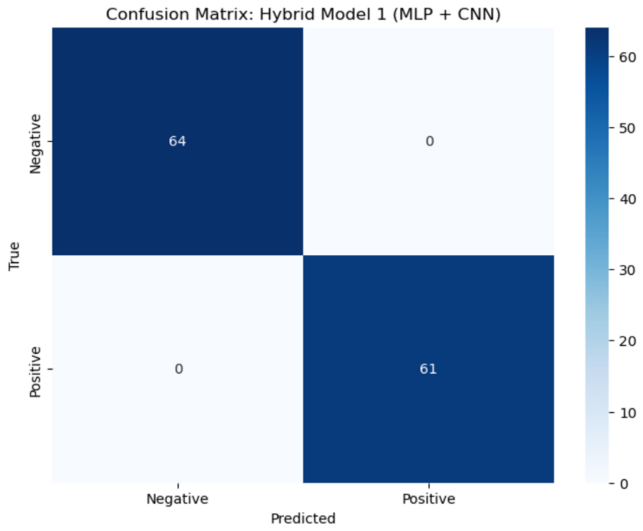

Fig. 7. Confusion matrix for Hybrid Model-1 (MLP-CNN).

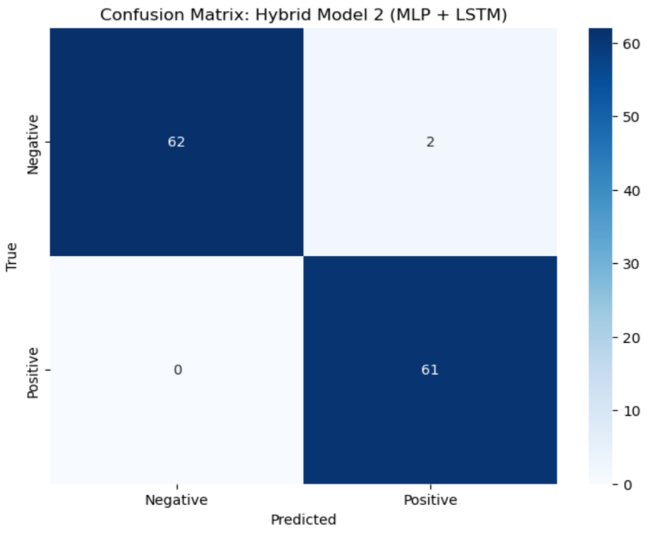

Fig. 8. Confusion matrix for Hybrid Model -2 (MLP-LSTM).

V. CONCLUSIONS

This study showcases that the proposed hybrid deep learning framework integrating feature selection with a Convolutional Neural Network (CNN)-Bidirectional Long Short-Term Memory (BiLSTM) architecture is highly effective for predicting novel drug-disease associations. The proposed model achieved an excellent performance, with an Area Under the Curve (AUC) of 0.957 by effectively capturing both spatial hierarchies and long-range sequential dependencies within complex biomedical data. The strategic use of Mutual Information (MI) and Recursive Feature Elimination (RFE) proved critical for enhancing interpretability and reducing noise from high-dimensional feature spaces. These results provide a robust, scalable, and biologically-informed computational strategy for accelerating drug repositioning efforts. Future work will focus on incorporating graph neural networks and multi-omics data fusion to further improve predictive comprehensiveness.

REFERENCES

[1] C. P. Adams and V. V. Brantner, "Estimating the Cost of New Drug Development: Is It Really $802 Million?," *Health Affairs*, vol. 25, no. 2, pp. 420–428, Mar. 2006, https://doi.org/10.1377/hlthaff.25.2.420.

[2] T. T. Ashburn and K. B. Thor, "Drug Repositioning: Identifying and Developing New Uses for Existing Drugs," *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 673–683, Aug. 2004, https://doi.org/10.1038/nrd1468.

[3] G. Cano *et al.*, "Automatic Selection of Molecular Descriptors Using Random Forest: Application to Drug Discovery," *Expert Systems with Applications*, vol. 72, pp. 151–159, Apr. 2017, https://doi.org/10.1016/j.eswa.2016.12.008.

[4] F. Napolitano *et al.*, "Drug Repositioning: A Machine-Learning Approach Through Data Integration," *Journal of Cheminformatics*, vol. 5, no. 1, Dec. 2013, Art. no. 30, https://doi.org/10.1186/1758-2946-5-30.

[5] H. Chen and J. Li, "A Flexible and Robust Multi-Source Learning Algorithm for Drug Repositioning," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, Boston, MA, USA, Aug. 2017, pp. 510–515, https://doi.org/10.1145/3107411.3107473.

[6] X. Liang *et al.*, "Predict and Interpret Drug–Disease Associations Based on Data Integration Using Sparse Subspace Learning," *Bioinformatics*, vol. 33, no. 8, pp. 1187–1196, Apr. 2017, https://doi.org/10.1093/bioinformatics/btw770.

[7] H. Liu, Y. Song, J. Guan, L. Luo, and Z. Zhuang, "Inferring New Indications for Approved Drugs via Random Walk on Drug-Disease Heterogenous Networks," *BMC Bioinformatics*, vol. 17, no. S17, Dec. 2016, Art. no. 539, https://doi.org/10.1186/s12859-016-1336-7.

[8] Q. Chen *et al.*, "ILDMSF: Inferring Associations Between Long Non-Coding RNA and Disease Based on Multi-Similarity Fusion," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 1106–1112, May 2021, https://doi.org/10.1109/TCBB.2019.2936476.

[9] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, "Computational Drug Repositioning Using Low-Rank Matrix Approximation and Randomized Algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1904–1912, Jun. 2018, https://doi.org/10.1093/bioinformatics/bty013.

[10] L. Yu, J. Huang, Z. Ma, J. Zhang, Y. Zou, and L. Gao, "Inferring Drug-Disease Associations Based on Known Protein Complexes," *BMC Medical Genomics*, vol. 8, no. S2, Dec. 2015, Art. no. S2, https://doi.org/10.1186/1755-8794-8-S2-S2.

[11] V. Martínez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "DrugNet: Network-Based Drug–Disease Prioritization by Integrating Heterogeneous Data," *Artificial Intelligence in Medicine*, vol. 63, no. 1, pp. 41–49, Jan. 2015, https://doi.org/10.1016/j.artmed.2014.11.003.

[12] H. Luo *et al.*, "Drug Repositioning Based on Comprehensive Similarity Measures and Bi-Random Walk Algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, Sep. 2016, https://doi.org/10.1093/bioinformatics/btw228.

[13] G. Fu, J. Wang, C. Domeniconi, and G. Yu, "Matrix Factorization-Based Data Fusion for the Prediction of lncRNA–Disease Associations," *Bioinformatics*, vol. 34, no. 9, pp. 1529–1537, May 2018, https://doi.org/10.1093/bioinformatics/btx794.

[14] W. Lan *et al.*, "LDICDL: LncRNA-Disease Association Identification Based on Collaborative Deep Learning," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 3, pp. 1715–1723, May 2022, https://doi.org/10.1109/TCBB.2020.3034910.

[15] Y. Liu, M. Wu, C. Miao, P. Zhao, and X.-L. Li, "Neighbourhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction," *PLOS Computational Biology*, vol. 12, no. 2, Feb. 2016, Art. no. e1004760, https://doi.org/10.1371/journal.pcbi.1004760.

[16] W. Zhang, Y. Chen, and D. Li, "Drug-Target Interaction Prediction through Label Propagation with Linear Neighbourhood Information," *Molecules*, vol. 22, no. 12, Nov. 2017, Art. no. 2056, https://doi.org/10.3390/molecules22122056.

[17] W. Zhang *et al.*, "Predicting Drug-Disease Associations by Using Similarity Constrained Matrix Factorization," *BMC Bioinformatics*, vol. 19, no. 1, Dec. 2018, Art. no. 233, https://doi.org/10.1186/s12859-018-2220-4.

[18] T. R. Noviandy, G. M. Idroes, A. Maulana, R. P. F. Afidh, and R. Idroes, "Optimizing Hepatitis C Virus Inhibitor Identification with LightGBM and Tree-Structured Parzen Estimator Sampling," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18810–18817, Dec. 2024, https://doi.org/10.48084/etasr.8947.

[19] J. Piñero *et al.*, "DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants," *Nucleic Acids Research*, vol. 45, no. D1, pp. 833–839, Jan. 2017, https://doi.org/10.1093/nar/gkw943.

[20] A. P. Davis *et al.*, "Comparative Toxicogenomics Database (CTD): Update 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. 1138–1143, Jan. 2021, https://doi.org/10.1093/nar/gkaa891.

[21] D. S. Wishart *et al.*, "DrugBank 5.0: A Major Update to the DrugBank Database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. 1074–1082, Jan. 2018, https://doi.org/10.1093/nar/gkx1037.

[22] L. M. Schriml *et al.*, "Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion," *Nucleic Acids Research*, vol. 47, no. D1, pp. 955–962, Jan. 2019, https://doi.org/10.1093/nar/gky1032.

[23] B. Jassal *et al.*, "The Reactome Pathway Knowledgebase," *Nucleic Acids Research*, Nov. 2019, Art. no. gkz1031, https://doi.org/10.1093/nar/gkz1031.

[24] J.-Y. Shi, A.-Q. Zhang, S.-W. Zhang, K.-T. Mao, and S.-M. Yiu, "A Unified Solution for Different Scenarios of Predicting Drug-Target Interactions via Triple Matrix Factorization," *BMC Systems Biology*, vol. 12, no. S9, pp. 498–503, Dec. 2018, https://doi.org/10.1186/s12918-018-0663-x.