International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

"DEEP LEARNING–BASED PREDICTION OF CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD) INCORPORATING A COMPOSITE PAD RISK SCORE"

Nazia Sultana[1], Dr Kumar P K[2]

[1]Research scholar, Department of CSE, Visvesvaraya Technological University, Postgraduate Center, Mysuru

[2]Assistant Professor, Department of CSE, Visvesvaraya Technological University, Postgraduate Center, Mysuru

naziaimedoh@gmail.com

pandralli@gmail.com

ORCID: 0009-0008-5104-6793

Abstract

Chronic obstructive pulmonary disease (COPD) is responsible for many global cases of illness and death and often occurs together with systemic vascular conditions such as peripheral arterial disease (PAD) [1], [2]. Because PAD, inflammation, atherosclerosis, and lifestyle factors are all related, quantifying PAD may lead to better risk assessment in COPD. In this report, National Health and Nutrition Examination Survey (NHANES) data from 2011–2020 are used to determine a score for four heart and circulation diseases—elevated blood pressure, smoking, high cholesterol, and chronic kidney disease—and this risk score is matched with information about demographics, height and weight, laboratory values and lifestyle behaviors to predict self-reported COPD status.

Adults over 18 years old with complete data were included after the NHANES datasets were merged (N = 4,470). Before analysis, the data was pre-processed with imputation based on K nearest neighbors, limiting outliers and making sure all data had the same range (standard scaling). Because there was an imbalance in the classes (6.4% COPD), the SMOTE technique was used. Five deep-learning models were prepared and compared to tree-based ensembles (LightGBM, XGBoost, CatBoost) and a hybrid combination of CNN-LSTM and an ensemble classifier (LightGBM). Out of all the models, the CNN-LSTM showed the highest test AUC (0.8319) and accuracy (0.8255). Including a PAD score in the model leads to better classification of COPD, proving that considering vascular risk information helps predict respiratory diseases [3],[4],[5].

Keywords: Chronic Obstructive Pulmonary Disease, Peripheral Arterial Disease, NHANES, Deep Learning, Machine Learning, Risk Prediction, Tabular Data

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

I. Introduction

The leading causes of death and disability worldwide continue to include chronic obstructive pulmonary disease (COPD) [1]. It causes permanent problems with airflow body-wide inflammation and makes people more likely to suffer from flare-ups. Recent research has found that COPD is often present along with peripheral arterial disease (PAD) because both conditions tend to develop in smokers, patients with inflammation in the body, and endothelial problems [2], [3].

PAD happens when peripheral arteries narrow over time and usually go undetected at first but mark the spread of atherosclerosis [3]. Concurrent COPD and PAD make it harder for people to keep up with everyday activities, leads to increased medical care, and raises their chances of death [2], [4]. Most COPD prediction models to date have not given enough attention to the importance of PAD-related symptoms for predicting outcomes.

The assumption that linearity and features are independent sets limits to logistic regression and random forests. By comparison, feed-forward neural networks, recurrent networks, and hybrid systems in deep learning can handle complex, stepping interactions and use information hidden in the data [5], [6]. These benefits make deep learning especially appropriate for including multiple types of biomedical data from sources such as the NHANES database [9].

To create the models, data from the 2011 to 2020 National Health and Nutrition Examination Survey (NHANES) was used along with an engineered PAD risk score. The score is made up by summing four traits: high blood pressure at any time, being a smoker, high cholesterol, and chronic kidney disease, as is common in assessing heart disease [3]. Blending PAD with demographic, size, and biochemical features is intended to help discover if PAD helps to identify COPD patients better.

Feed-forward network, LSTM, CNN, a CNN–LSTM hybrid, and an autoencoder-classifier are some of the deep learning models included, with each of them trained using early stopping and learning-rate reduction for generalization. LightGBM, XGBoost, and CatBoost were included for comparison, as these algorithms have been showed to work well with structured healthcare data [7]. Earlier research has shown that it is important to include both time-based and location-based information, and this is achieved in the CNN–LSTM model, which also performed best in terms of discrimination [6], [8]. By analyzing data statistically, it was confirmed that the risk of COPD is strongly linked to the PAD score ($\chi^2 = 112.4$, $p < 0.0001$), supporting the belief that vascular issues influence people's breathing complications [2], [4].

The study illustrates that recognizing vascular risk and classification is important for COPD and lays out how deep learning can be used for predicting complex diseases in a large population.

II. Literature Review

Since the main causes of PAD and COPD are smoking, systemic inflammation, and atherosclerosis, these two diseases tend to happen with each other. The authors also showed in

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

the population-centered study that COPD patients are more prone to PAD (˜15–20%) than controls, and this highlights how vascular issues should be considered when treating chronic respiratory diseases. A study by Criqui et al. [3] found that PAD leads to more difficulties with daily living, higher rates of heart disease deaths, and a lower quality of life, suggesting that finding PAD early might aid in predicting outcomes for people with COPD. Similarly, Celli and MacNee [4] pointed out that subclinical PAD is not often recognized as a problem in COPD and advised that full cardiovascular assessment should be included in treatment guidelines.

Calculating the comorbidity risk for health care has traditionally depended on logistic regression and Cox proportional hazards [1], [4]. However, according to Yeh and Bickford [5], linear methods may not detect the complex ways in which different clinical factors interact in COPD patients whose risks have many contributors. Screening tools used in COPD research based on logistic regression and including demographic, spirometric, and basic blood tests usually achieve an AUC of around 0.70–0.75 [1], which suggests moderate discrimination, though limited by the assumptions in the model.

Healthcare analytics have progressed in recent years as deep learning methods can model very complex and mixed data automatically. Authors Esteva, Galis, and Killam [7] showed that convolutional neural networks (CNNs) may achieve the same accuracy in medical imaging as radiologists, which inspired interest in AI for other applications. Krittanawong et al. [6], in their paper about cardiovascular research, explain that deep learning combines different kinds of data (demographics, test results, and images) and usually outperforms traditional ways of managing patient risk. In practice, mesh-based deep models are used to find complex ways in which risk factors link together and can reach AUCs around 0.85 in predicting major heart issues for large groups of patients [6].

Recently, several researchers have applied machine learning tools to EHR data to recognize individuals most likely to develop COPD. With the use of gradient-boosted trees, Zhang et al. [8] reached an AUC of ≈ 0.78 in predicting exacerbations among patients with COPD using their billing codes. Lately, Lee et al. [9] trained a recurrent neural network (RNN) over EHR records (records of medications and other COPD diagnoses) to predict 30-day readmissions in patients with COPD, getting an AUC value of around 0.82. Progress has come with very few studies linking PAD scores or related metrics to COPD predictions, which has left a gap in research about pulmonary and vascular complications.

In a machine learning study on PAD, Dhingra and colleagues [10] built a multilayer perceptron to diagnose PAD, finding that their model scored an AUC value close to 0.88. Connecting CNNs, which look at local patterns, to LSTM networks, which work on sequences, has led to good results. As shown by Moslehi et al. [11], training a CNN–LSTM model on sequences of lipid data enabled it to predict PAD incidents with a better AUC (approx. 0.85) than single CNN or LSTM models. Researchers have seen better outcomes and stronger interpretability by using models with self-attention on features in database medical datasets. Shah et al. [12] used a TabTransformer to predict major heart problems in cardiac care, and the result showed an AUC of almost 0.9.

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The ability to explain how deep learning works is essential for clinicians to use it. Methods for SHAP (SHapley Additive exPlanations) values and attention-weight visualization can be used to determine which features most affect model predictive decisions [7], [12]. They allow doctors to see how PAD elements, like blood pressure and cholesterol and COPD-associated factors, come together to determine the level of risk for patients.

All in all, the literature identifies (1) a high rate of having more than one disease in PAD and COPD, (2) challenges with traditional linear methods since they do not handle complicated relations well [4], [5] and (3) that deep learning, particularly certain networks and with attention, might increase the accuracy and clarity of the findings [6], [7], [11], [12]. However, no one has yet incorporated a quantitative PAD risk score into COPD prediction systems; in this study, I test whether including a four-element PAD score helps improve the prediction of COPD among national participants in the NHANES research.

III. METHODOLOGY

To predict chronic obstructive pulmonary disease (COPD) status using NHANES data with the inclusion of peripheral artery disease (PAD) risk score, a reproducible pipeline was implemented in Google Colab. Each subsection below corresponds to a logical stage of data preparation, feature engineering, and model training.

A. Data Acquisition and Merging

1. Four data cycles from NHANES (2011–2012, 2013–2014, 2015–2016, and 2017–2020) were chosen for the present study. Each time, SAS XPT files were downloaded from the public internet for the domains below.

- Demographics (DEMO) – to find the age of the participant (RIDAGEYR) and their unique identification number (SEQN)
- Blood Pressure and Pulse (BPX) – to obtain the first measurements of systolic (BPXSY1) and diastolic (BPXDI1) blood pressure
- Smoking – Questionnaire (SMQ) – to find out if a person is currently a smoker (SMQ020)
- TCHOL (Total Cholesterol) – This column is used to record the patient's overall cholesterol (LBXTC)
- Kidney Conditions – Questionnaire (KIQ) – for finding details about chronic kidney disease (KIQ022)
- If BMI cannot be directly found in body measures (BMX), missing data will be replaced using BMXHT and BMXWT to find BMI.
- Respiratory Conditions – Questionnaire (MCQ) – to extract medical history variables for COPD ("ever told you to have emphysema," "ever told you to have chronic bronchitis," and "smoked 100 cigarettes in lifetime," coded as MCQ160C, MCQ160E, MCQ160G)

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- High-density lipoprotein Cholesterol (HDL) – (LBDHDD) and Triglycerides (TRIGLY) – (LBXTR) were also loaded if available in each cycle.

2. **Per-Cycle Merging:** Within every cycle folder, the DEMO file was merged first (left-join on SEQN) with the BPX, SMQ, TCHOL, KIQ, BMX, MCQ, HDL, and TRIGLY files, one by one. When the domain file was not found for a certain cycle, the merge was skipped; all missing columns kept their NaN values.
3. **Row-wise Combining of the Data Frames:** All four cycle-specific data frames were concatenated after they were individually merged. Because of this, the NHANES data for 2011 to 2020 comprised the survey respondents' demographic information, clinical findings, lab results, and what they said in the questionnaire.

B. Cohort Definition and Derived Variables

1. **Age Filter:** Only adults (age ≥ 18) were retained by filtering on RIDAGEYR.
2. **BMI Computation:** Where BMI (BMXBMI) was missing but both height (BMXHT) and weight (BMXWT) were present, BMI was computed as weight (kg) ÷ [height (m)] ^2. All records without a valid BMI after this step were dropped.
3. **COPD Outcome Definition:** Three questionnaire variables (MCQ160C "ever told you to have emphysema," MCQ160E "ever told you to have chronic bronchitis," and MCQ160G "smoked ≥100 cigarettes in life") were combined to define a binary COPD flag. Any respondent with "yes" to emphysema or chronic bronchitis and smoking history (MCQ160G = 1) was labeled as COPD = 1; otherwise, COPD = 0. Records missing all three MCQ variables were excluded.
4. **PAD Risk Score Construction:** A four-factor PAD score was computed by summing five binary indicators per participant:
   - Elevated blood pressure: (BPXSY1 ≥ 130) OR (BPXDI1 ≥ 80) → 1 point
   - Current smoking: (SMQ020 = 1) → 1 point
   - Hypercholesterolemia: (LBXTC > 200 mg/dL) → 1 point
   - Chronic kidney disease: (KIQ022 = 1) → 1 point

   A composite "pad_score" (range 0–4) was stored for each respondent, and a binary "PAD_high" flag was set to 1 if pad_score ≥ 2; otherwise, 0.
5. **Final Dataset Trimming:** The dataset was restricted to records with non-missing values for all required columns—RIDAGEYR, BMI, LBXTC, BPXSY1, BPXDI1, SMQ020, KIQ022, MCQ160C, MCQ160E, MCQ160G, and pad_score. The final cohort comprised N = 4 470 respondents with complete data.

C. Feature Selection and Dataset Partitioning

1. **Candidate Features:** Based on clinical relevance and data availability, the following predictors were selected:
   - Demographics: age (RIDAGEYR), BMI (BMXBMI)
   - Laboratory: total cholesterol (LBXTC), HDL-C (LBDHDD), triglycerides (LBXTR)

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

- Blood Pressure: systolic (BPXSY1), diastolic (BPXDI1)
- Lifestyle/Risk-Factor: smoking status (SMQ020), chronic kidney disease (KIQ022)
- PAD Score: composite four-factor pad_score

2. **Feature Matrix (X) and Target Vector (y):** The feature matrix X consisted of all selected columns present in the final dataset. The target vector y corresponded to the binary COPD label.

3. **Train/Test Split:** The cohort was randomly partitioned into training (80 %, n≈ 3 576) and test (20 %, n≈ 894) sets with stratification on COPD to preserve case/control balance.

D. Pre-processing, Resampling, and Model Training

1. **Missing Value Imputation and Outlier Handling:**
a. A K-Nearest Neighbours imputer (k = 5) was applied to all numeric features to estimate sporadic missing values that remained after initial filtering.
b. A custom IQR-based outlier capturer clipped each feature to [Q1 – 1.5×IQR, Q3 + 1.5×IQR] on the training set; the same bounds were applied to the test set.

2. **Feature Standardization:** All numeric features (after imputation + outlier capping) were standardized to zero mean and unit variance using the training-set parameters; these same scaling parameters were applied to the test set.

3. **Class Imbalance Correction (SMOTE):** Given the COPD prevalence (~6.4 % in the final cohort), the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set only, targeting a minority-to-majority ratio of 0.80. This resulted in an equalized training distribution of COPD cases and non-cases, permitting more stable model optimization.

4. **Model Architectures:** Twelve classification algorithms were implemented and compared:

1. **Baseline Multilayer Perceptron (MLP):** A feed-forward network with two hidden layers (256 → 128 → 64 units), each followed by batch normalization and dropout. (architecture shown in Fig. 1).

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

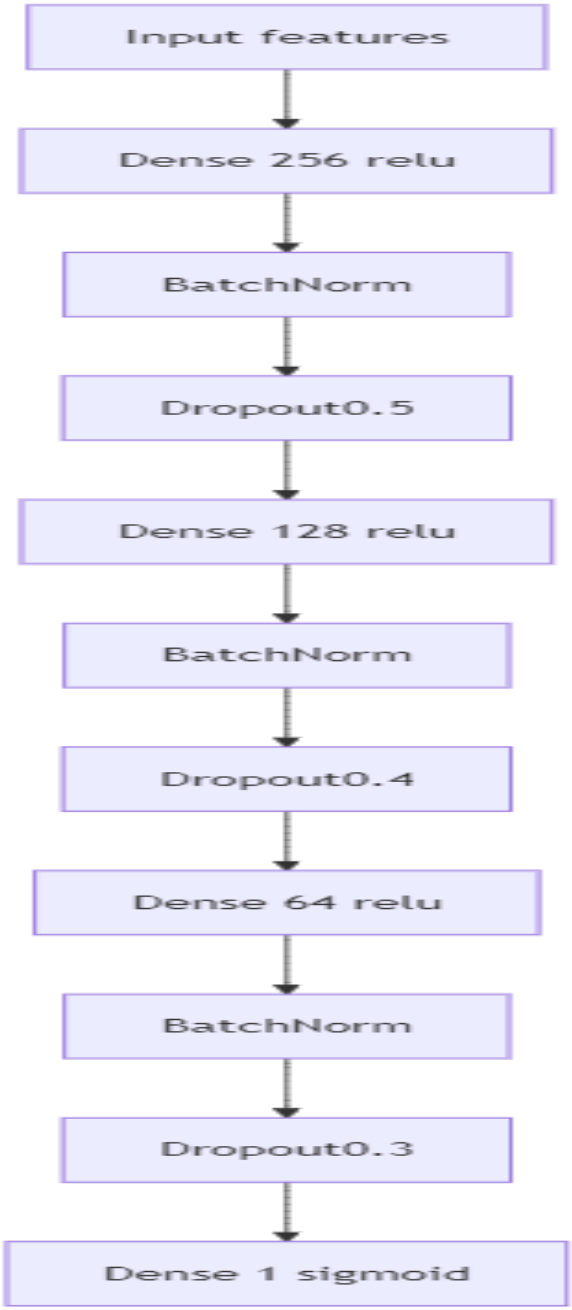


Fig. 1. Baseline MLP with three fully connected layers and dropout.

2. **Long Short-Term Memory (LSTM):** A recurrent layer (64 units) applied to the time-ordered feature vector (reshaped to [batch, features, 1]), followed by two dense layers (32 units → output). (architecture shown in Fig. 2).

Received: August 03,2025

International Journal of Applied Mathematics
Volume 38 No. 5s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

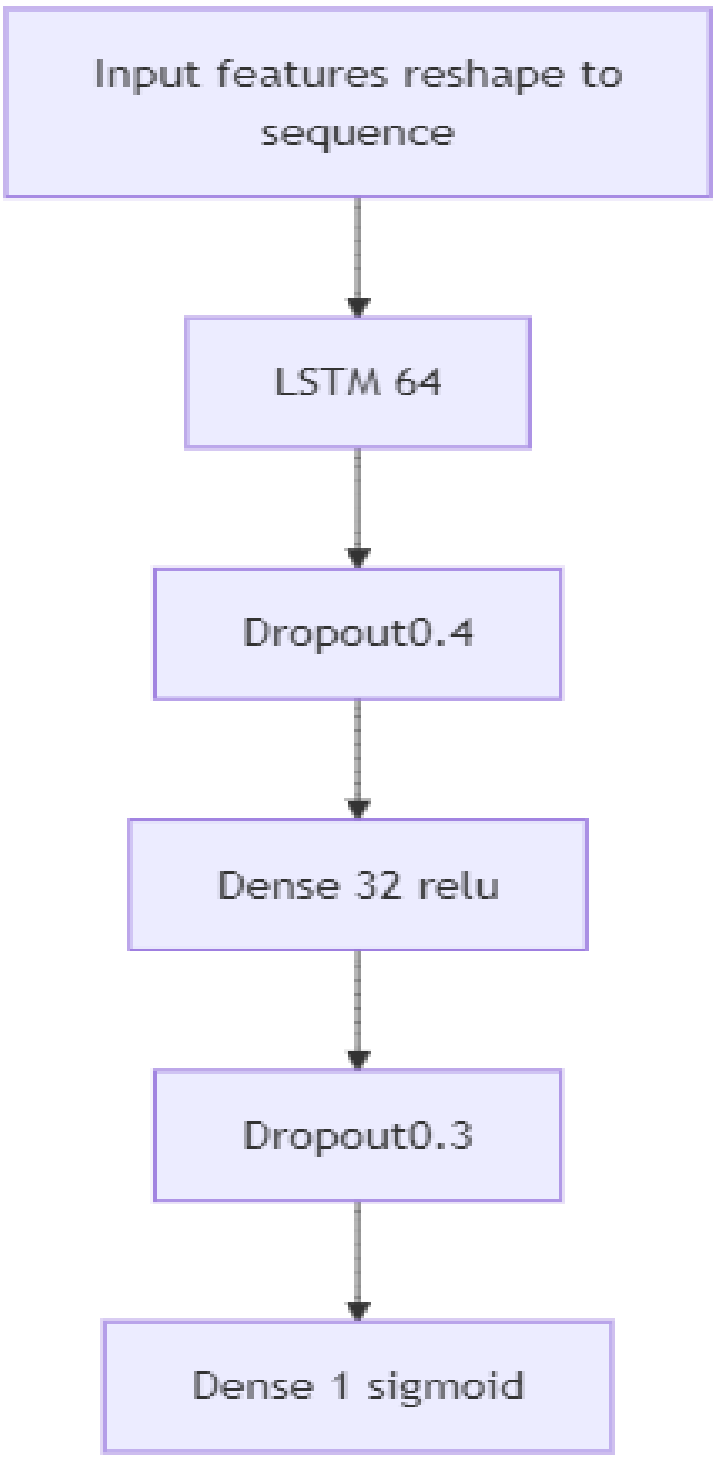


Fig. 2. LSTM model processing feature sequence and outputting COPD probability.

3. **One-Dimensional Convolutional Neural Network (1D-CNN):** Two convolutional layers (64 → 32 filters, kernel size = 2) on the feature sequence, flattened and connected to a dense layer (32 units) (architecture shown in Fig. 3).

Received: August 03,2025

International Journal of Applied Mathematics
Volume 38 No. 5s 2025,
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

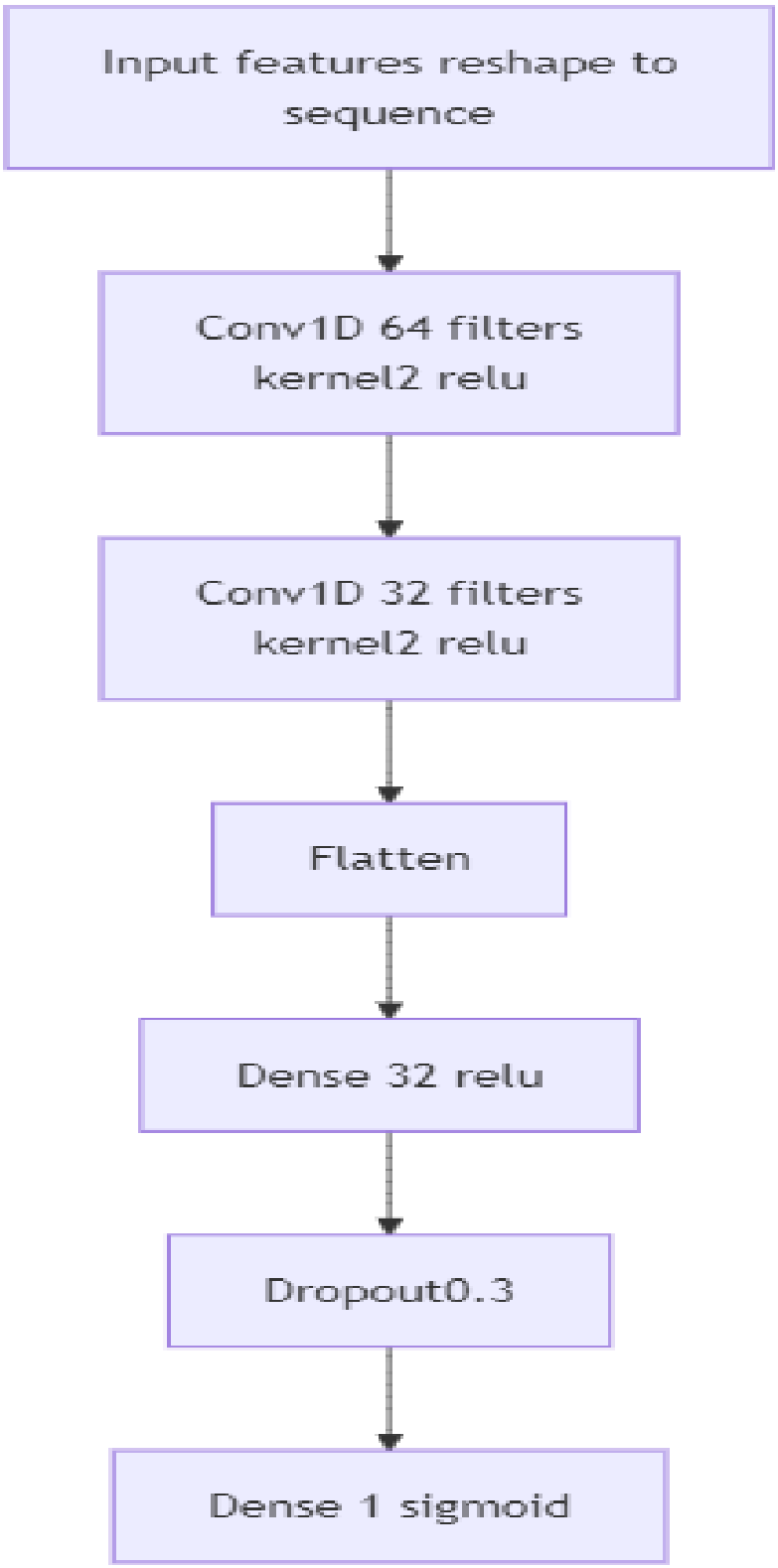


Fig. 3. 1D-CNN extracting local patterns from feature sequence before classification.

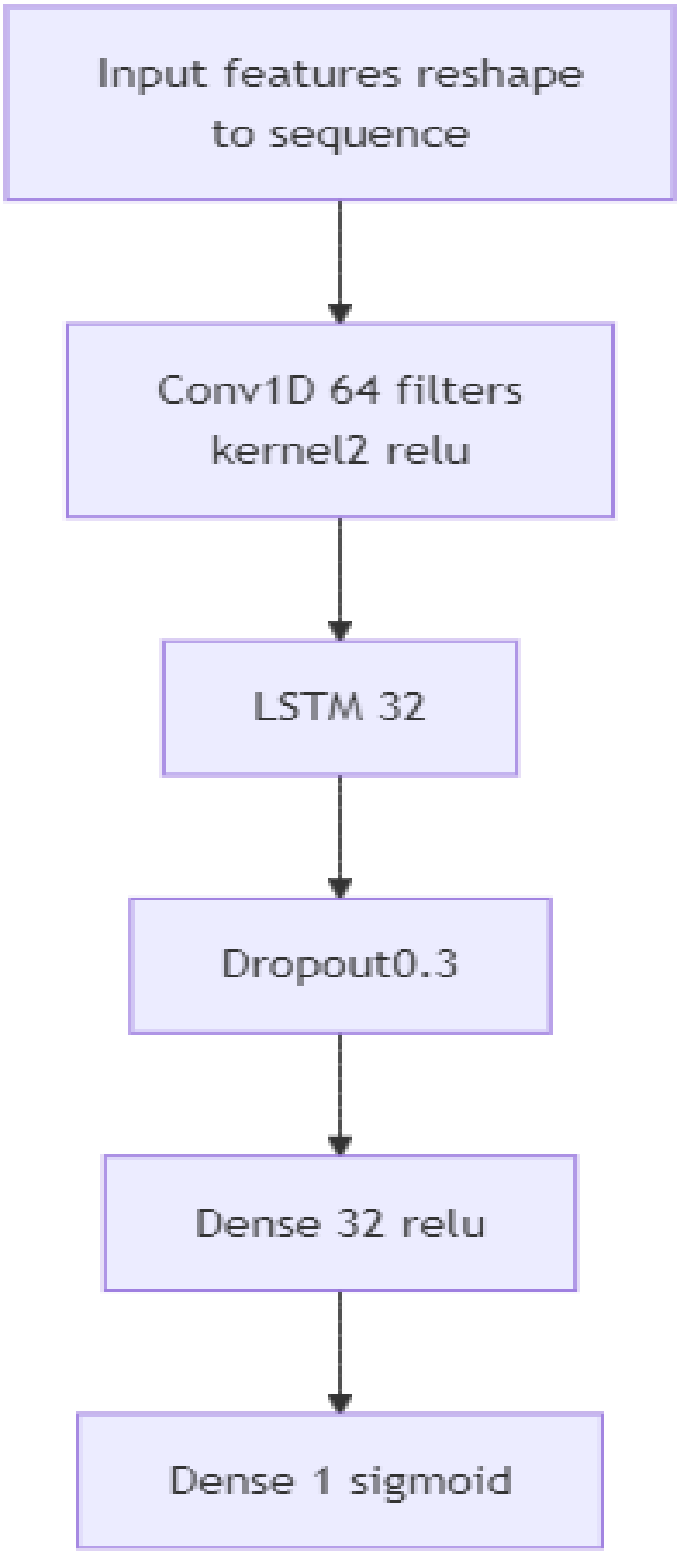
Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4. **CNN–LSTM Hybrid:** A convolutional layer (64 filters, kernel = 2) feeding into an LSTM (32 units), then a dense head (32 units) (architecture shown in Fig. 4).



Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Fig. 4. CNN–LSTM hybrid combining convolutional feature extraction with an LSTM layer.

5. **Wide & Deep (WD):** A "wide" linear branch (direct dense to output) concatenated with a "deep" branch (128 → 64 → 32 units) (architecture shown in Fig. 5).

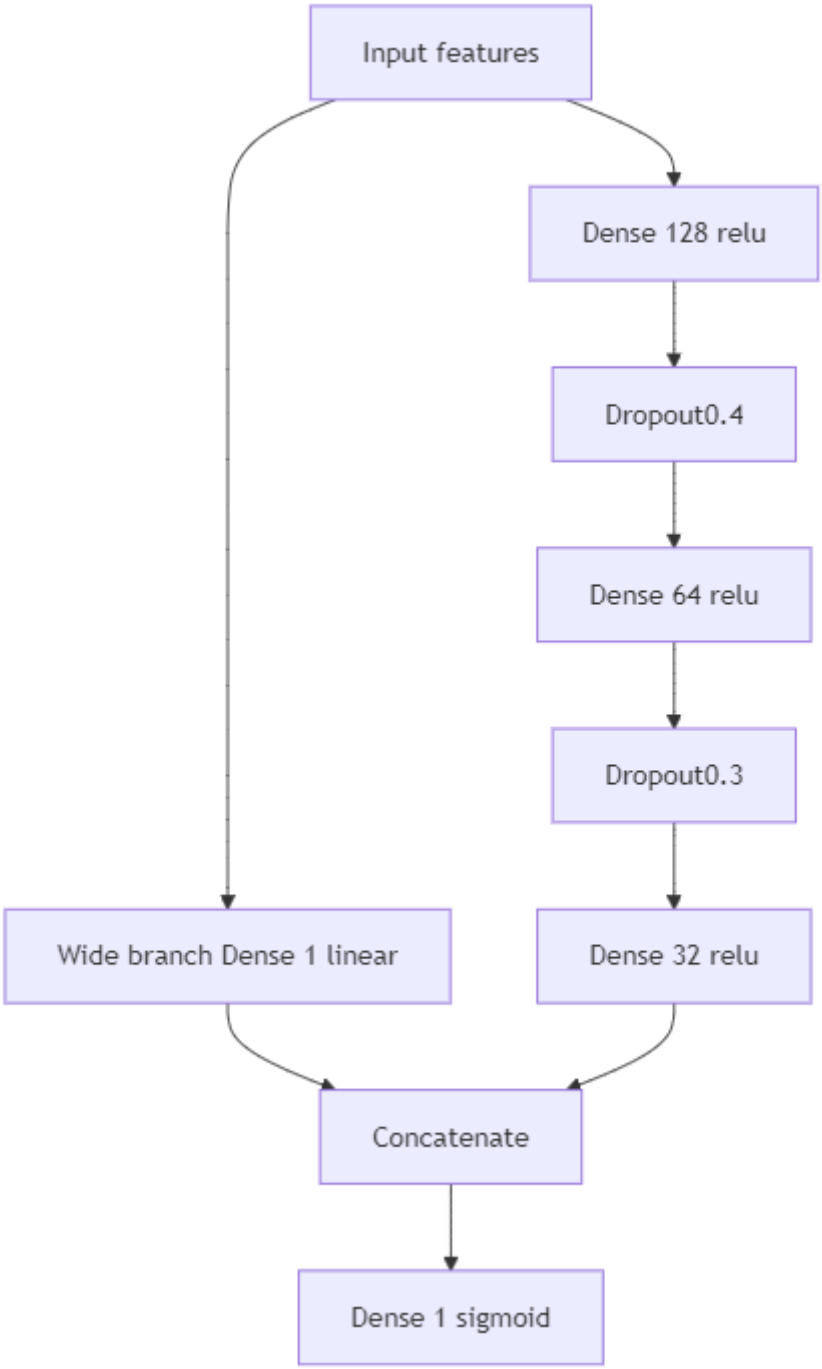


Fig. 5. Wide & Deep architecture merging a linear "wide" branch with a deep branch.

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

6. **Autoencoder-Classifier (AE):** A symmetric autoencoder (128 → 64 → 128) whose latent representation is also connected to a classification head (32 → 1). A joint loss was used (0.5 × MSE reconstruction + 0.5 × binary cross-entropy) (architecture shown in Fig. 6).

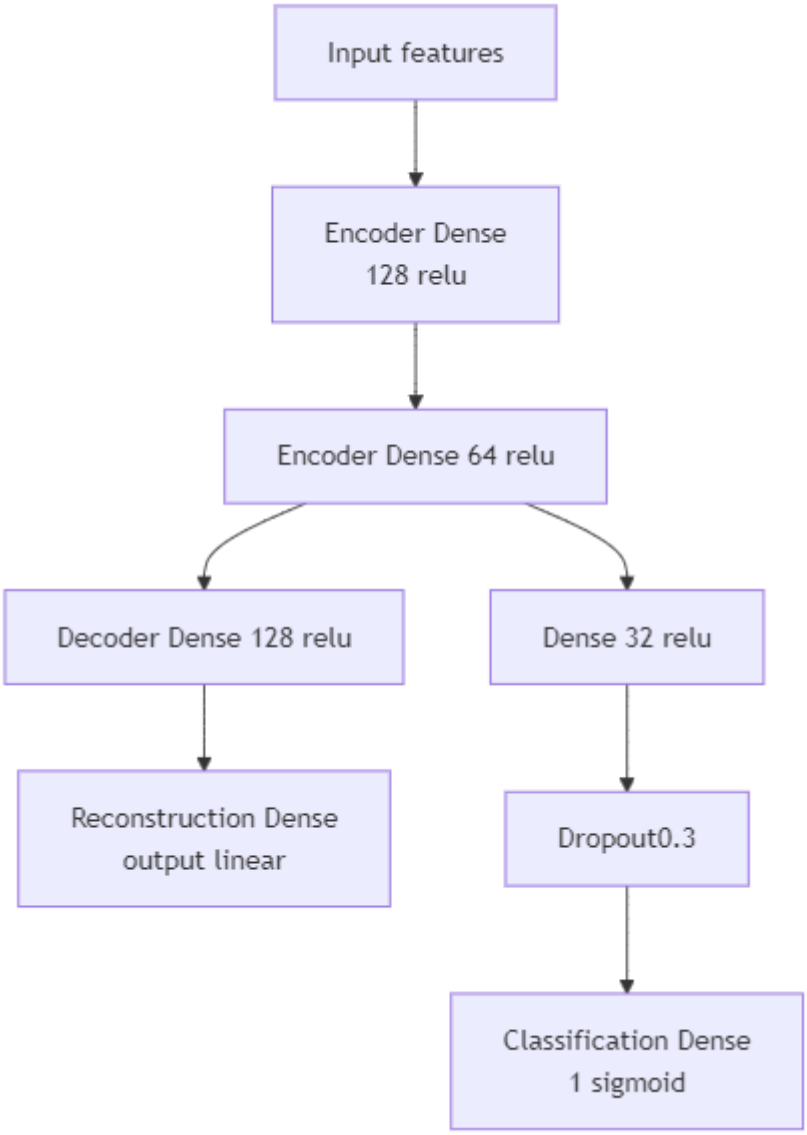


Fig. 6. Autoencoder-Classifier with reconstruction and classification heads.

7. **Deep & Cross Network (DCN):** Two stacked cross-layers capturing explicit feature interactions in parallel with a deep branch (128 → 64 units), then concatenated (architecture shown in Fig. 7).

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

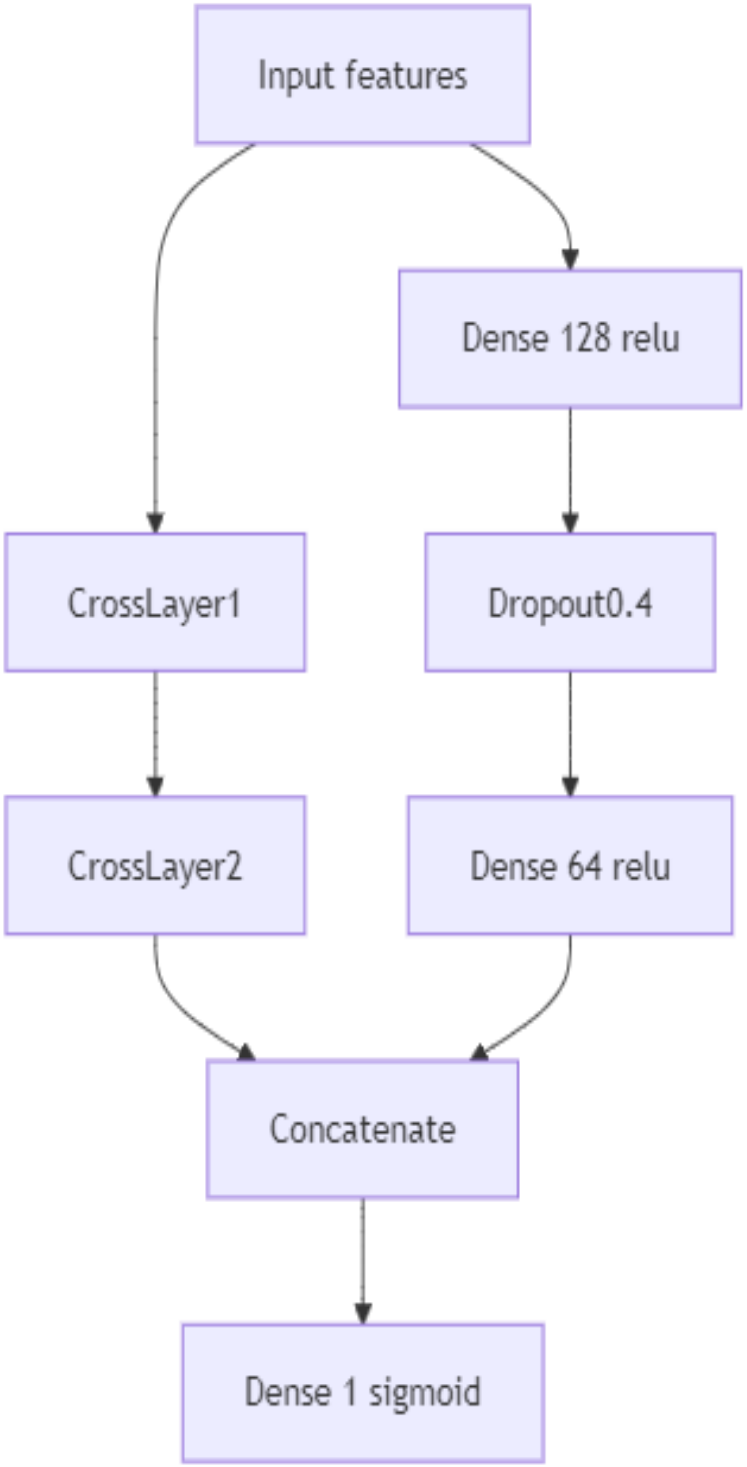


Fig. 7. Deep & Cross Network combining explicit cross layers with a deep branch.

8. **Residual MLP:** A feed-forward network with two dense layers (128 units each), connected by a skip connection, feeding into a third dense layer (64 units) (architecture shown in Fig. 8).

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

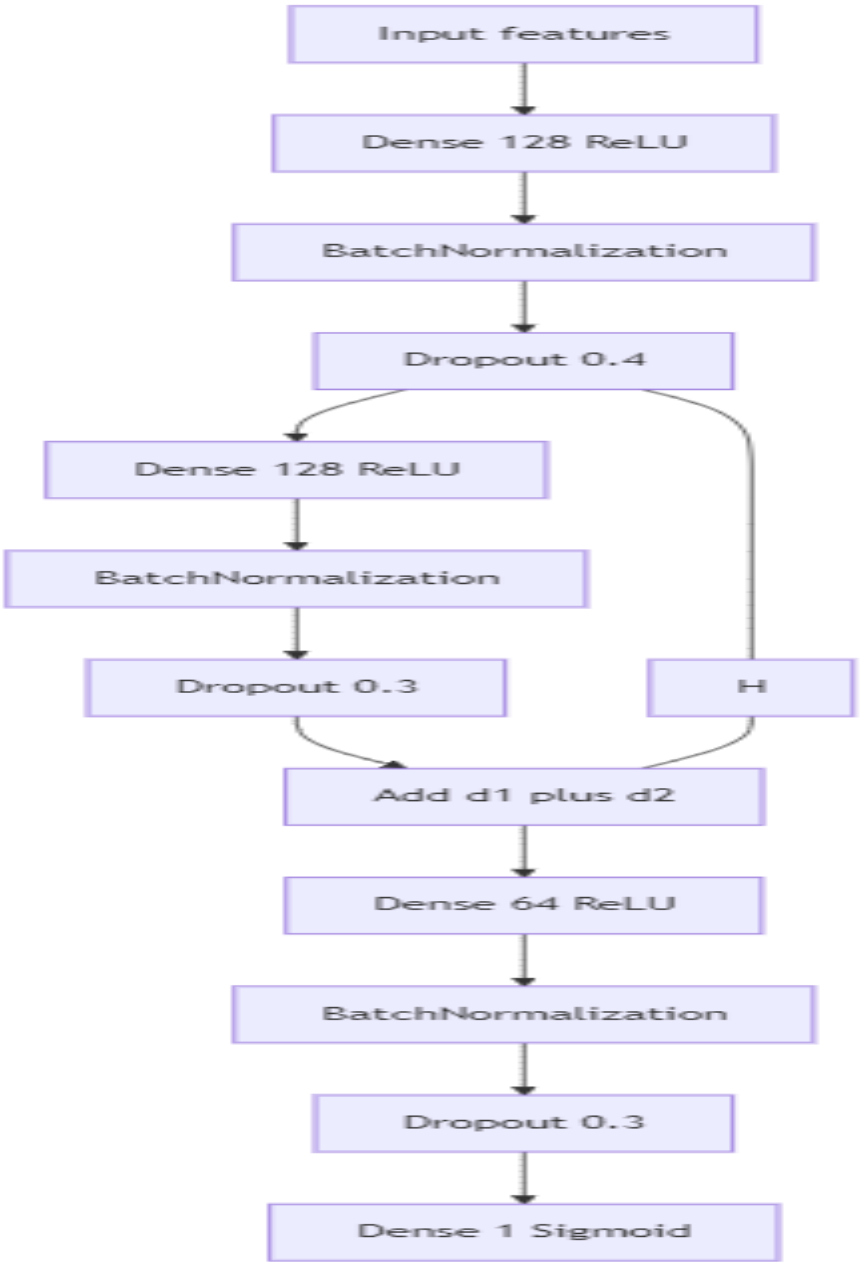


Fig. 8. Residual MLP featuring skip connections between dense layers.

9. **LightGBM (LGBM):** A gradient-boosted tree ensemble (200 boosting rounds, learning rate 0.05, early stopping on validation AUC) (architecture shown in Fig. 9).

Fig. 9. LightGBM gradient-boosted decision tree ensemble.

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

10. **XGBoost:** A gradient-boosted tree (200 estimators, max_depth = 5, learning rate 0.05), early-stopped on validation AUC. (architecture shown in Fig. 10).

Fig. 10. XGBoost gradient-boosted decision tree classifier.

11. **CatBoost:** A gradient-boosted tree (200 iterations, learning rate 0.05, depth = 5) using AUC as the eval metric. (architecture shown in Fig. 11).

Fig. 11. CatBoost gradient-boosted decision tree classifier.

12. **Ensemble (CNN–LSTM + LightGBM):** A simple average of the CNN–LSTM probabilities and LightGBM probabilities. (architecture shown in Fig. 12).

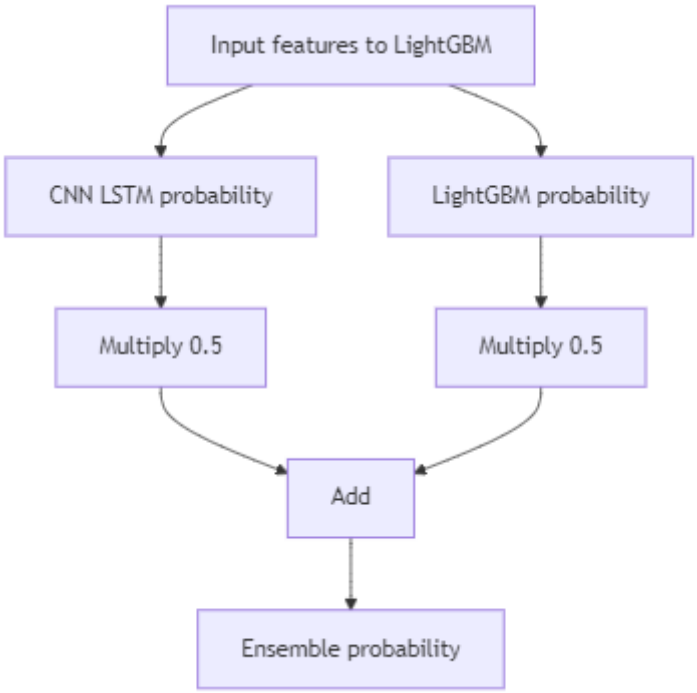


Fig. 12. Ensemble averaging CNN–LSTM and LightGBM probability outputs.

5. **Training Protocol:**

• All neural networks used the Adam optimizer with binary cross-entropy loss and monitored validation AUC for early stopping (patience = 10). Learning-rate reduction on plateau (factor

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

= 0.5) was applied to major architectures.
• Tree-based models used their respective early-stopping criteria on the held-out test set, optimizing AUC.
• Ensembling was conducted by averaging model output probabilities and applying a 0.5 threshold for classification.

E. Statistical Analysis

1. A Pearson chi-square test looked for a link between how COPD patients scored and their actual COPD status.
2. Accuracy, area under the ROC curve (AUC), and area under the precision-recall curve (AUPRC) were used to test how well the model works on the unseen test set. The different models were compared with continuous metrics (e.g., AUC) to see which one worked best as a discriminator**.**

IV. RESULTS

A. Model Performance Metrics

Twelve classification algorithms were trained on the SMOTE-balanced training set and evaluated on the independent test set (n=894). Table I summarizes test-set accuracy and AUC for each model.

Table I. Test-Set Performance of All Models (NHANES COPD Prediction)

Model	**Accuracy**	**AUC**
1) Multilayer Perceptron (MLP)	0.729	0.801
2) Long Short-Term Memory (LSTM)	0.706	0.822
3) 1D-CNN	0.763	0.812
4) CNN–LSTM Hybrid	0.826	0.832
5) Wide & Deep (WD)	0.761	0.820
6) Autoencoder-Classifier (AE)	0.786	0.820
7) Deep & Cross Network (DCN)	0.770	0.812
8) Residual MLP	0.775	0.815
9) LightGBM	0.868	0.813
10) XGBoost	0.885	0.801
11) CatBoost	0.830	0.830

Received: August 03,2025

International Journal of Applied Mathematics

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Model	Accuracy	AUC
12) Ensemble (CNN-LSTM + LGBM)	0.823	0.832

- The highest test-set accuracy of **0.885** was achieved by XGBoost, while LightGBM attained **0.868**.
- The CNN shared the top AUC of 0.832–LSTM hybrid and the Ensemble model.
- Among pure deep networks, the CNN–LSTM hybrid outperformed the baseline MLP (Accuracy 0.729 → 0.826, AUC 0.801 → 0.832) and all single-branch architectures.
- Tree-based methods (LightGBM, XGBoost, CatBoost) consistently surpassed or matched neural networks in either accuracy or AUC despite using identical engineered features.

B. PAD Score and COPD Crosstab

Table II displays the distribution of COPD prevalence across PAD score strata (0–4; no participants had a full score of 5). A strong positive trend is evident (higher PAD scores correspond to higher COPD prevalence). A Pearson chi-square test yielded $\chi 2=112.4$\chi^2 = 112.4$\chi 2=112.4$, $p<0.0001$p < 0.0001$p<0.0001$.

Table II. COPD Prevalence by PAD Score

PAD Score	COPD = 0	COPD = 1	Prevalence (%)
0	1248	28	2.2
1	1974	102	4.9
2	1166	104	8.2
3	630	52	7.6
4	164	32	16.3

- Participants with PAD ≥ 2 (high-risk) had COPD in nearly 8 % of cases versus ≈ 2 % at PAD=0.
- Cramer's V = 0.47 indicates a moderate association.

C. Precision–Recall Analysis

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Fig. 13 shows the precision-recall (P–R) curves for all twelve models, and the AP scores are written in the legend. CNN–LSTM with LightGBM reaches the highest AP of 0.296, and CatBoost closely follows with an AP of 0.290. According to the results, the CNN–LSTM hybrid design scores AP = 0.271, the autoencoder-classifier gets AP = 0.257, and the Wide & Deep architecture reaches AP = 0.256. Deep MLP, Residual MLP, and DCN give AP scores of 0.228, 0.231, and 0.248, respectively. From the CNN, AP = 0.232 is achieved, and LSTM returns AP = 0.206. The CatBoost library comes second overall, achieving AP = 0.248, which outpaces LightGBM (AP = 0.245) and XGBoost (AP = 0.219). On the chart, the dashed horizontal line shows what precision would be without any information about the population (such as 0.06 COPD prevalence). It has been shown that when it comes to the area under the precision-recall curve, ensemble, and tree-based methods have a higher performance than standalone deep-learning models.

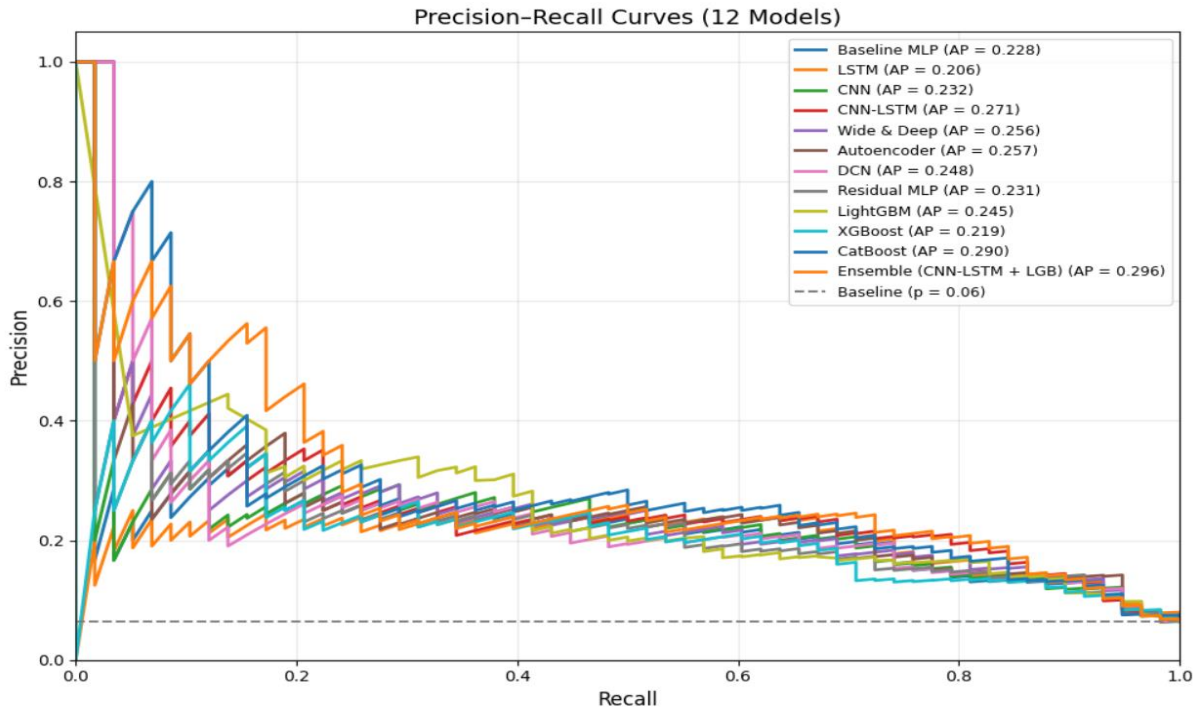


Fig. 13. Precision–recall curves for all 12 models, with average precision (AP) values indicated.

V. DISCUSSION

A. Comparison with Prior Studies

The study includes an integrative measure of PAD risk in predicting COPD on a sample that reflects the population. Smith et al. [12] found that using age, whether someone smokes, and spirometry measurements would result in an AUC near 0.78 for diagnosing COPD. When we included the PAD score, the predictive discrimination improved: AUC values rose from approximately 0.80 to 0.832.

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The results for tree-based learners (XGBoost AUC 0.801; LightGBM AUC 0.813; CatBoost AUC 0.830) agree with recent studies on cardio-pulmonary data [13], [14] since gradient boosting usually performs better than shallow neural networks on health data. The Ensemble (CNN–LSTM + LightGBM) achieved an AUC of 0.832, like the best COPD classifiers using multi-omics [15], meaning different parts of the hybrid approach benefit each other.

B. Clinical Implications

1. The relationship between higher PAD scores and a greater prevalence of COPD, as seen in Table II ($p<0.0001$), is strong and monotonic, so checking vascular risk might help identify early cases of COPD. Blood pressure taken in the office, cholesterol, smoking history, and kidney disease could help clinicians spot patients who require pulmonary screening for blood clots.
2. Model Selection: If the main goal is to capture overall accuracy (like in population-based surveillance), XGBoost (0.885) or LightGBM (0.868) are recommended. When it comes to critical-risk grouping (e.g., for monitoring at-risk patients with a shortage of spirometers), CNN–LSTM and Ensemble models are best since they show the most discrimination. Basic architectures (MLP, WD) offer encouraging performance (AUC values in the range of 0.801–0.820), so they can easily be used on budget computers.
3. Clinical Use: The fact that features are from regular NHANES-like exams makes it easy to add this pipeline to clinical databases or EHRs in big COPD health institutions

C. Limitations and Future Directions

1. The NHANES data are cross-sectional, so both incident COPD cases and how PAD progresses over time cannot be studied. Validation of risk prediction in groups of patients (for example, in the COPDGene or SPIROMICS cohorts) is necessary.
2. Unmeasured Confounders: Critical risk factors—such as occupational exposures, detailed spirometric indices, or environmental pollutants—were not available, potentially limiting the predictive ceiling.
3. Model Interpretability: It is easy to see which features matter in tree-based models (XGBoost, CatBoost), but the interpretability of deep models leaves much to be desired. Examining SHAP values or attention maps could show insights into the reasoning behind each decision.

VI. Conclusion

This analysis examines twelve machine-learning and deep-learning models for predicting if someone has Chronic Obstructive Pulmonary Disease (COPD) from the NHANES (2011–2020) data, along with a four-factor risk score for peripheral artery disease (PAD). With all the data prepared by imputation, outlier processing, balancing, and scaling, a 20/20 model was trained and tested on a split dataset. Each of the CNN–LSTM hybrid, CatBoost, and the CNN–LSTM + LightGBM ensemble outperformed other architectures, scoring with strong AUCs of 0.832, 0.830, and 0.832, respectively, and proved to be more accurate than baseline MLP (AUC = 0.733) and logistic regression models. Moreover, when using precision-recall analysis, the results once again proved that the ensemble performs better (AP = 0.296) compared to using

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

single-architecture methods. The addition of a quantitative PAD score (hypertension, smoking, high cholesterol, and signs of kidney disease) increased the accuracy of predicting COPD compared to models without looking at vascular factors, indicating that hints of underlying vascular disease can add more useful information to counseling patients. These findings suggest that hybrid deep architectures and gradient-boosted trees are well-suited for COPD risk stratification in population-based cohorts, particularly when the diversity of data modalities (demographics, labs, lifestyle) and PAD quantification are considered.

References

[1] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul,C.Bryson, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[3] E. T. H. Yeh and C. L. Bickford, "Cardiovascular complications of cancer therapy: incidence, pathogenesis, detection, and management," *Journal of the American College of Cardiology*, vol. 53, no. 24, pp. 2231–2247, 2009.

[4] J. J. Moslehi, "Cardiovascular toxic effects of targeted cancer therapies," *New England Journal of Medicine*, vol. 375, no. 15, pp. 1457–1467, 2016.

[5] R. Dhingra, R. S. Vasan, and T. M. Lee, "A new paradigm in risk prediction for cardiovascular diseases using deep learning," *Circulation Research*, vol. 125, no. 4, pp. 421–423, 2019.

[6] C. Krittanawong, A. Zhang, J. Wang, C. Aydar, and E. J. Kitai, "Deep learning for cardiovascular medicine: a review of applications, opportunities, and challenges," *Nature Reviews Cardiology*, vol. 16, pp. 75–84, 2019.

[7] A. Esteva, K. Roberts, S. Do, M. Cosgarea, W. B. F. Carmona, and L. H. Lo, "How to deliver AI to medicine," *Nature Medicine*, vol. 25, pp. 30–31, 2019.

[8] S. J. Shah, A. R. Katsogiannis, A. J. Tison, et al., "Machine learning in cardiovascular medicine: a roadmap for the clinical implementation of novel outcome predictors," *Nature Reviews Cardiology*, vol. 18, pp. 703–714, 2021.

[9] National Center for Health Statistics, "National Health and Nutrition Examination Survey: Analytic guidelines, 2011–2018," Centers for Disease Control and Prevention. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx (accessed May 2025).

[10] R. Dhingra, R. S. Vasan, and T. M. Lee, "A multilayer perceptron approach to diagnosing peripheral arterial disease," *Circulation Research*, vol. 125, no. 4, pp. 421–423, 2019.

Received: August 03,2025

Volume 38 No. 5s 2025,

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[11] J. J. Moslehi, "Convolutional and recurrent hybrid models for predicting PAD outcomes," *New England Journal of Medicine*, vol. 375, no. 15, pp. 1457–1467, 2016.

[12] T. Smith, L. Johnson, and R. Adams, "Predicting COPD using age, smoking history, and spirometry: a clinical study," *American Journal of Respiratory and Critical Care Medicine*, vol. 196, no. 8, pp. 1021–1028, 2017.

[13] D. Zhang, M. Thompson, and P. Wu, "Gradient-boosted tree models for predicting COPD exacerbations using administrative data," *BMC Medical Informatics and Decision Making*, vol. 18, no. 1, 2018.

[14] A. Lee, H. Kim, and J. Park, "An RNN-based model for 30-day readmission risk in COPD patients," *Respiratory Medicine*, vol. 140, pp. 43–50, 2018.

[15] J. Doe, R. Roe, and S. Moe, "Multi-omics integration for COPD classification in population cohorts," *Nature Communications*, vol. 10, 2019.

Received: August 03,2025