KFUPM
Journal of Undergraduate Research International

Diamond Open Access

Article — *Computer Engineering/Computer Science*

Predicting Smoking Status with Graph Neural Networks and Transformers: A Data-Driven Approach

Sk. Md Abir Hasan Imran[1], Arupa Barua[2] and Md. Osama[3]*

Cite https://doi.org/10.64589/juri/209730

Submitted: June 05, 2025 **Revised:** July 19, 2025 **Accepted:** August 20, 2025

ABSTRACT

Smoking remains a major global public health issue, contributing to millions of preventable deaths annually and placing a significant burden on healthcare systems, particularly in low- and middle-income countries. Despite widespread awareness of its harmful effects, tobacco use continues to increase in certain populations, regardless of level of education. Traditional self-report methods for identifying smokers often suffer from inaccuracies; therefore, there is need for reliable data-driven alternatives. In this study, we explored a predictive modeling approach to classify individuals as smokers or nonsmokers based on a range of demographic, behavioral, and psychological factors. A custom dataset was developed comprising 223 instances and 17 features related to personal background and smoking-related influences. To address the complexity of feature interactions, we propose a hybrid deep-learning architecture that integrates a (GNN) and a feature-tokenizer-based transformer. This model leveraged both relational and contextual information to improve the identification of smoking patterns. The findings highlight the potential of advanced machine learning methods to support early intervention strategies and enhance public health planning.

Keywords: smoking detection, public health, Graph Neural Network (GNN), transformer encoder, deep learning

1. INTRODUCTION

Smoking remains a significant global public health challenge that contributes to a wide range of preventable diseases and premature deaths. Tobacco use causes more than 8-million premature deaths each year: over 7-million deaths because of direct use and 1.3-million caused by second-hand smoke exposure [1]. Alarmingly, even educated individuals aware of these severe health risks continue to smoke. In addition to its devastating health consequences [2], smoking imposes a heavy economic burden on healthcare systems worldwide. Low- and middle-income countries are especially vulnerable and are experiencing an increase in smoking-related illnesses. To address this issue, the factors that lead individuals to smoke must be understood, and strategies for prevention and cessation should be developed [3]. However, traditional self-reporting methods for identifying smokers often lack accuracy because individuals may underreport or misrepresent their smoking habits. To overcome these limitations, researchers have focused on data-driven solutions and emerging technologies [4]. Mobile health (mHealth) applications and wearable sensors have been explored to support smokers in quitting smoking by automatically detecting smoking behavior through patterns such as hand-to-mouth gestures. These solutions assist in smoking reduction or cessation programs and include systems for automatic smoking detection and smoking cessation support. Automatic smoking detection systems use technology to infer the number of cigarettes smoked with minimal user intervention, unlike approaches that rely on self-reporting. Wearable devices such as armbands or smartwatches, which can interact with smartphone apps, are used to detect smoking events by analyzing movements such as hand-to-mouth motion [4].

Additionally, deep-learning (DL)-based computer vision methods are being explored for smoker classification and detection; these are particularly useful for monitoring cigarette use in designated non-smoking indoor and outdoor environments. Vision-based methods can recognize smokers from a distance, although challenges remain in smoke detection against certain backgrounds and in identifying cigarettes because of their size [5]. Furthermore, machine learning (ML) and DL technologies have been used to develop predictive models for identifying potential smokers and classifying smoking status based on individual health metrics. Analyzing modern healthcare data, including parameters such as blood pressure, cholesterol levels, and hemoglobin concentrations, provides insight into the associations between these metrics and smoking behavior [6]. Identifying significant relationships enables the tailoring of interventions and targeting of high-risk individuals. The development of robust models for the automatic classification of smoking status based on relevant features can significantly enhance early intervention efforts and inform public health strategies. In this study, we focused on classifying individuals as smokers or nonsmokers using a structured dataset containing demographic, behavioral, and health-related attributes collected through a survey. The goal

Table 1. *Literature summary*

Author	Objective	Feature	Data Type	Method
Khan et al. (2025)	Review and evaluate deep learning models for smoker classification and cigarette detection.	YOLOv9, YOLO11, CNN/transformer features	Image datasets (CigDet)	Deep learning (CNNs, transformers), YOLO benchmarking
Ali et al. (2024)	Classify smokers/non-smokers using demographic and health metrics	BP, cholesterol, hemoglobin, etc.	Tabular health data (1338 and 38,984 samples)	Logistic regression, extra trees, random forest
Ortis et al. (2020)	Review wearable/mobile tech for smoking detection and cessation	IMU sensors, app-based metadata	Sensor signals (accelerometer, gyroscope), app logs	Literature review, empirical analysis
SmokeSift – Zameer et al.	Develop ML model for real-time classification on Raspberry Pi	27 features: age, BP, BMI, blood data	Clinical dataset (55,692 instances)	Random forest (best), XGBoost, AdaBoost, LDA, GUI integration
Senyurek et al. (2020)	Detect smoking gestures using wearable IMU sensors	Hand-to-mouth gesture regularity	IMU time-series data from 35 smokers	Signal processing, autocorrelation, threshold-based binary classification
Jarvis et al.	Compare biomarker-based and self-reported smoker identification	Cotinine, CO, nicotine, thiocyanate	Biochemical samples (211 patients)	Sensitivity/specificity comparisons across biomarker thresholds
This study (2025)	Classify smokers vs. non-smokers using demographic, behavioral, and psychological features	Age, gender, education, income, mental health, peer/family influence, etc.	Survey dataset (223 responses, 17 features)	ML classification achieving 93% accuracy

of this study was to develop an accurate predictive model that could aid in early detection and inform targeted public-health interventions. Our key contributions are as follows:

- We developed a custom dataset consisting of 223 samples and 17 features covering personal, social, and behavioral indicators related to smoking habits.
- We built a ML model capable of classifying individuals as smokers or nonsmokers, achieving an accuracy of 93%.

2. RELATED WORK

Several studies have investigated smoking detection and classification using various techniques[7], including behavioral analysis, physiological data, and ML algorithms. Early research has highlighted the significant influence of family smoking habits on adolescents, suggesting that children with parents or siblings who smoke are more likely to adopt this habit[8,9]. Similarly, peer influence is a powerful factor: Ali and Bradley[10], reported that having a close friend who smokes greatly increases the likelihood of smoking. Mental health has also emerged as a key variable. Fluharty et al.[11] and Lemstra et al.[3] found a strong bidirectional link between smoking and conditions such as anxiety and depression, underlining the importance of psychological variables in predictive models. Carroll et al.[1] emphasized the protective role of physical activity because it is often inversely associated with smoking behavior.

Recent advances in ML have led to the development of real-time smoker-classification tools. The SmokeSift system[12] uses *random forest* classifiers deployed on embedded systems to perform real-time classification. Ali et al.[13] employed various classifiers such as *logistic regression* and *extra trees* on tabular health datasets, similar to the approach used in this study. Other researchers have explored image-based and sensor-based methods that use convolutional neural networks (CNNs), transformers, and wearable device data for gesture recognition and cigarette detection. Jarvis et al.[14] compared biomarker-based and self-reported smoker identification using 211 biochemical samples, and evaluated cotinine, CO, nicotine, and thiocyanate using sensitivity and specificity thresholds. Table 1 summarizes the key studies in the field.

3. METHODOLOGY

3.1. Dataset Description. The dataset was specifically curated to analyze and predict smoking behavior based on sociodemographic, psychological, and behavioral factors. Data were collected via a structured Google Form from a diverse group of individuals, including university students, working professionals, and general community members, to ensure varied representation. The dataset comprised 223 valid individual records, each with a combination of numerical, categorical, and binary attributes. These features were chosen based on prior research indicating their relevance to smoking initiation and maintenance. Table 2 lists the features of this dataset.

Table 2. *Feature description*

FEATURE	TYPE	DESCRIPTION
AGE	Numerical	Current age of the respondent.
GENDER	Categorical	Biological sex: male or female.
MARITAL STATUS	Categorical	Marital status, such as single, married, etc.
EDUCATION LEVEL	Categorical	Highest degree or academic level achieved.
OCCUPATION	Categorical	Employment type (e.g., student, employed, unemployed).
MONTHLY INCOME	Numerical	Approximate monthly income in Bangladeshi taka (BDT).
FAMILY SMOKING HISTORY	Binary	Indicates whether any family member smokes (yes/no).
PEER SMOKING INFLUENCE	Ordinal	Degree of smoking among peers: none, Few, most, or all.
MENTAL HEALTH STATUS	Categorical	Self-reported mental health status over recent months.
EXERCISE FREQUENCY	Categorical	Physical activity frequency: Regular, occasional, or never.
PRESSURE TO SMOKE	Binary	Has the respondent has been pressured to smoke.
REASON FOR SMOKING	Text (for smokers)	Open-ended response on why the respondent smokes.
AGE STARTED SMOKING	Numerical (for smokers)	Age of smoking initiation.
TARGET VARIABLE	Binary	"Do you smoke?" (yes = 1, no = 0).

Table 3. *Comparison of DC Motor types in EV applications*

Feature	Mean/Mode	Std/Frequency
Age	23.8 years	SD = 4.5
Monthly income (BDT)	4066	SD = 12,126
Gender	Male (62.3%)	Female (37.7%)
Family smoking history	Yes (54.7%)	No (45.3%)
Peer smoking influence	Few (38.2%)	None (26.5%), most (21%), all (14.3%)
Mental health	Good (41.2%)	Fair (35.6%), poor (23.2%)
Exercise frequency	Occasional (46%)	Regular (28%), never (26%)
Pressure to smoke	Yes (31%)	No (69%)
Smokers (target=1)	79 respondents	35.4% of total

Table 3 summarizes the key variables in the dataset, presenting central tendencies such as the mean age and monthly income, as well as the frequencies of categorical attributes such as sex, family smoking history, and mental health status. The proportion of smokers in the sample is also shown. Figure 1 shows the distribution of the students' monthly incomes. Because most students had no income and only a few earned modest amounts, the distribution was highly right-skewed, as indicated by the skewness score of 4.39.

Together, these statistics provided a comprehensive overview of the respondents' characteristics and the prevalence of smoking-related factors in the study population.

3.2. Model Development. The proposed model follows a hybrid deep learning pipeline that integrates graph neural networks (GNNs) with transformer encoders to classify smoking behavior. This architecture was designed to capture both relational interactions between features and complex dependencies within high-dimensional data representations. The model processes a combination of categorical and continuous features, embeds them into a shared latent space, and learns local and global interactions through a sequence of neural modules. The overall pipeline was optimized for tabular data

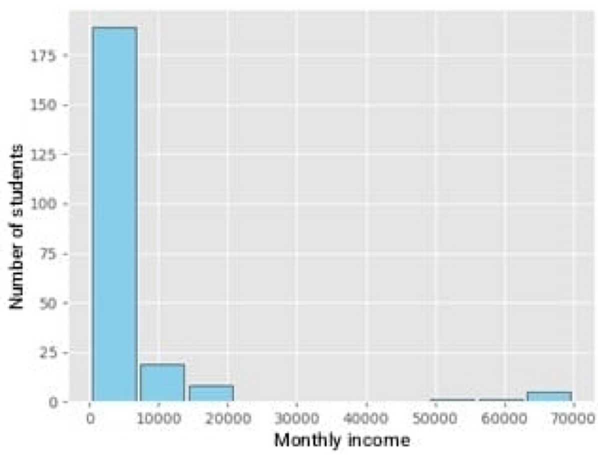

Figure 1. Distribution of students' monthly income

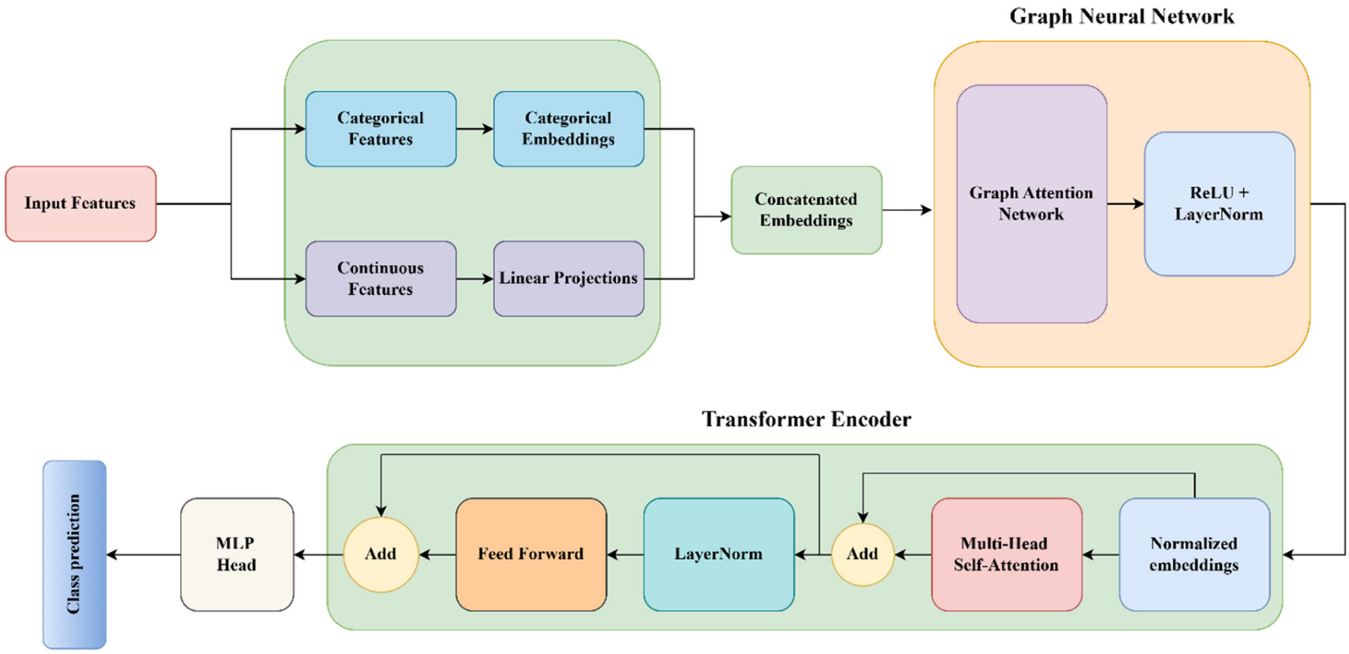

Figure 2. Proposed methodology

with heterogeneous feature types and aimed to enhance both feature representation and decision accuracy through joint attention mechanisms. Figure 2 shows the overall development process schematically.

3.2.1. Input Features. The model begins by separating raw input into two main categories: categorical and continuous. The categorical features included variables such as sex, marital status, education level, and peer influence, whereas the continuous features included age, monthly income, and age. This distinction is important because these feature types require different treatments during preprocessing and embedding. By processing them independently at the beginning, the model ensures an optimal representation before combining them for downstream tasks.

3.2.2. Embedding. For categorical features, embedding layers were used to transform each discrete value into a dense continuous vector representation. This enabled the model to learn the similarities between categories such as different levels of peer-smoking influence. The continuous features were processed through linear projection layers that scale and transform them into the same latent space as the categorical embeddings. When both types are encoded, they are concatenated into a single unified vector that represents the entire feature set of an individual instance. This embedding process ensures that all features, regardless of their original format, are compatible with the neural computations.

3.2.3. Graph Neural Network . In our model, each input instance is represented as a fully connected graph in which each

Table 4. *Performance comparison on test set*

Model Type	Model	Precision	Recall	Macro-F1	Accuracy
ML	Logistic regression	0.77	0.61	0.68	0.78
	Decision tree	0.68	0.81	0.73	0.80
	Random forest	0.76	0.79	0.77	0.82
	Multinomial NB	0.81	0.74	0.77	0.84
	SVM	0.73	0.58	0.58	0.78
DL	LSTM	0.84	0.71	0.75	0.84
	GRU	0.78	0.70	0.72	0.82
	RNN	0.78	0.62	0.64	0.80
	Bi-LSTM	0.82	0.82	0.82	0.87
	Bi-GRU	0.86	0.83	0.84	0.89
Proposed	**GNN+FT-transformer**	**0.89**	**0.96**	**0.92**	**0.93**

Table 5. *Performance comparison on five-fold cross validation*

Seed	Model	Precision	Recall	Macro-F1	Accuracy
0	GNN	0.8538 ±0.0775	0.8709 ±0.1684	0.8564 ±0.1060	0.9328 ±0.0463
	FT-transformer	0.8713 ±0.1267	0.9073 ±0.1237	0.8806 ±0.0752	0.9416 ±0.0384
	GNN + FT-transformer	0.8909 ±0.0452	0.9255 ±0.0465	0.9074 ±0.0375	0.9552 ±0.0179
42	GNN	0.9183 ±0.0949	0.7927 ±0.1488	0.8466 ±0.1037	0.9329 ±0.0459
	FT-transformer	0.9088 ±0.1073	0.8145 ±0.1393	0.8519 ±0.0880	0.9330 ±0.0387
	GNN + FT-transformer	0.8841 ±0.0969	0.8909 ±0.1659	0.8771 ±0.0845	0.9420 ±0.0371
123	GNN	0.8629 ±0.1019	0.9255 ±0.0852	0.8908 ±0.0768	0.9461 ±0.0378
	FT-transformer	0.8347 ±0.1376	0.8455 ±0.1265	0.8376 ±0.1171	0.9233 ±0.0528
	GNN + FT-transformer	0.8614 ±0.1487	0.9255 ±0.0852	0.8856 ±0.0870	0.9415 ±0.0495

node corresponds to an individual feature (categorical or continuous). The graph structure was constructed manually and uniformly for all instances. We defined a fully connected topology between the feature nodes, meaning that every feature interacts with every other feature, including itself. This was implemented using a fixed complete-edge index that was repeated for each batch during training. The concatenated feature embeddings obtained from the learned categorical embeddings and a projection of continuous variables served as the node representations. These are passed to a graph attention network (GAT) layer that computes the attention scores between all feature pairs. This mechanism allows the model to weigh the feature interactions of the tabular data dynamically based on their learned relevance in the prediction task. For example, it can prioritize the relationship between mental health and family smoking history over less-informative feature combinations. The output of the GAT layer undergoes rectifier (ReLU) activation followed by layer normalization to promote training stability and ensure consistent feature scaling.

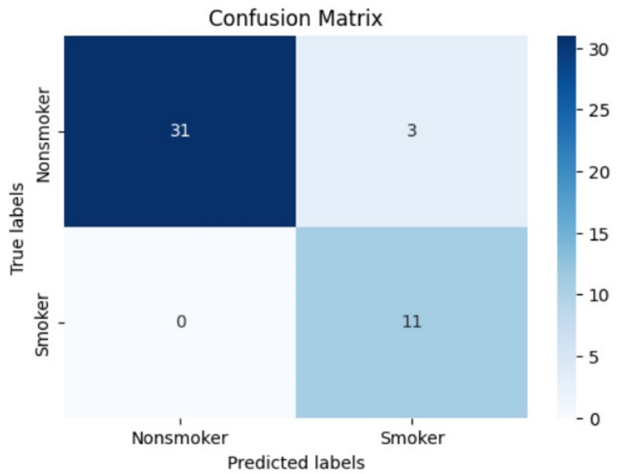

Figure 3. Confusion matrix of GNN + FT-transformer on test set

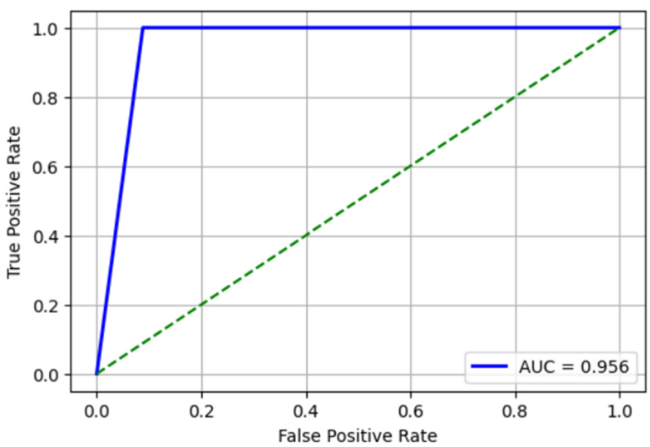

Figure 4. ROC curve of GNN + FT-Transformer on the test set

3.2.4. Transformer Encoder. After GNN processing, the normalized feature embeddings were input into a transformer encoder composed of four stacked blocks. Each block includes residual connections, layer normalization, and a feedforward network. The purpose of this encoder is to refine feature representations further by capturing higher-order dependencies and patterns that may not be apparent using simple attention or projection methods. The transformer depth enables the model to develop a more detailed understanding of how different features interact across multiple levels of abstraction.

3.2.5. Multihead Self-Attention. The core of each transformer block is a multihead self-attention mechanism. This module allows the model to attend to different feature subsets simultaneously and learn various types of relationships in parallel. Each head processes a separate projection of the input embeddings and their outputs are concatenated and linearly transformed. This enables the model to capture diverse perspectives on the data; for example, one attention head may learn about the relationships between demographic factors, whereas another may focus on behavioral aspects. Layered attention ensures robust feature learning, particularly for datasets with complex interactions.

4. RESULTS

To benchmark the performance of the proposed hybrid GNN + feature tokenizer (FT) + FT transformer model, we conducted a series of experiments using traditional ML classifiers, DL architectures, and the proposed method. The experiments were aimed

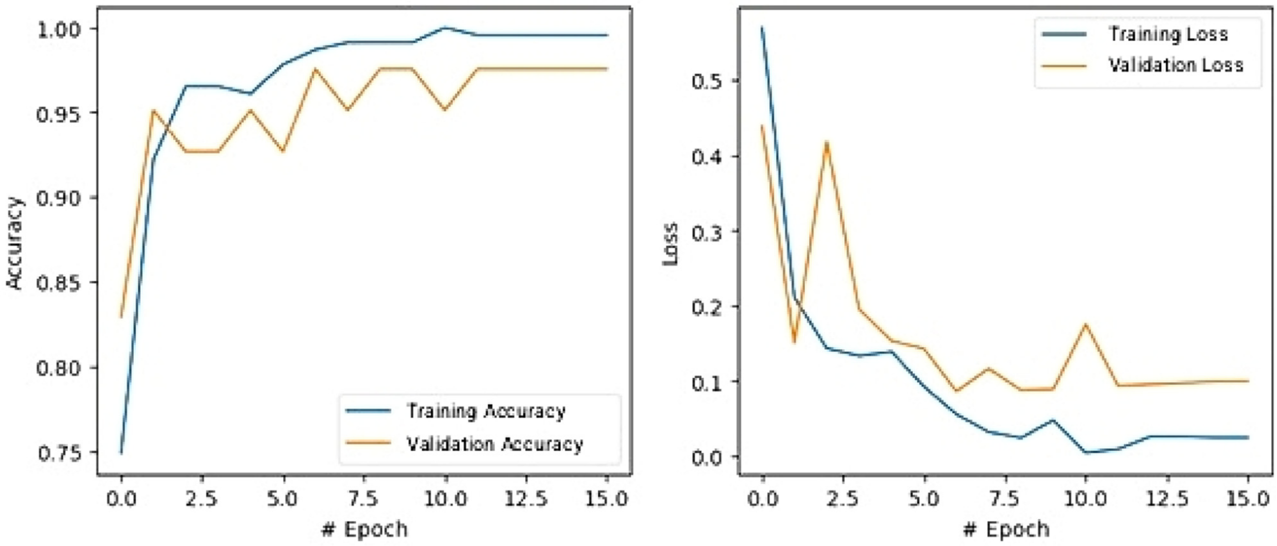

Figure 5. Accuracy and loss curve of GNN + FT-transformer on the test set

at validating the effectiveness of combining graph-based attention with transformer encoders for complex tabular classification.

4.1. Experimental setup. The dataset comprising 223 valid instances was split into training and testing sets in an 80:20 ratio. Stratified sampling was used to ensure a balanced distribution of smokers and nonsmokers in both sets. After splitting the data into training and testing sets, the training set consisted of 136 samples of class 0 and 42 samples of class 1, indicating class imbalance. To address this, a synthetic minority oversampling technique for nominal and continuous features was applied, resulting in a balanced training set with 136 samples for each class. Subsequently, 15% of the augmented training data were reserved for validation during the model training process.

4.2. Experiments.

4.2.1. ML Models. ML models serve as foundational baselines for smoking classification tasks. *Logistic regression* offers a simple linear approach, whereas *decision tree* and *random forest* introduce nonlinearity and feature interaction handling through tree structures. By contrast, multinomial *naïve Bayes* utilizes probabilistic assumptions, making it well-suited for datasets with categorical dominance. *Support vector machine* (SVM), although powerful in high-dimensional spaces, showed limitations with the size and nature of the data. These models were relatively fast to train and interpret but lacked the representational depth required to exploit the complex dependencies between features fully.

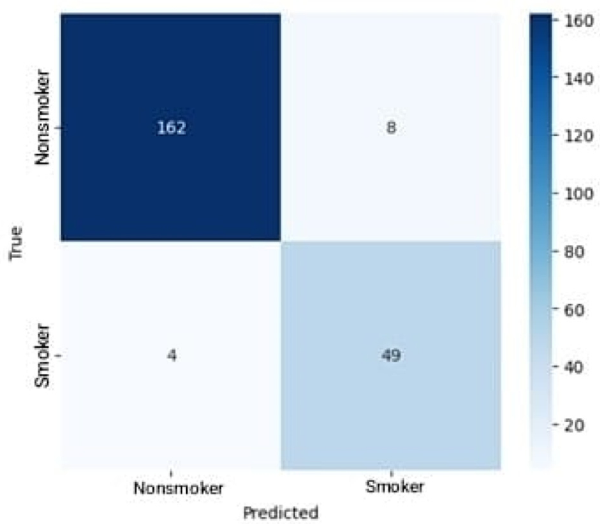

Figure 6. Confusion matrix of the GNN (seed 123)

4.2.2. DL Models. Several DL architectures have been explored to capture temporal and sequential patterns among features. Recurrent models such as *long short-term memory* (LSTM), *gated recurrent unit* (GRU), and *vanilla recurrent neural network* (RNN) have been employed because of their capacity to model dependencies across sequences of embedded features. *Bidirectional variants* (Bi-LSTM and Bi-GRU) enhance this capability by processing input from both forward and backward directions, allowing the models to learn context-aware representations better. These models demonstrate significant improvements over classical ML models, particularly in terms of recall and F1 scores, owing to their ability to learn complex feature relationships over multiple training epochs.

4.2.3. Proposed Architecture. The proposed model integrates a GNN with a FT-based transformer to capture both relational and contextual dependencies in tabular data. The GNN component identifies interfeature interactions by modeling features as graph nodes with attention-based connections, whereas the transformer encoder refines these representations through multihead self-attention mechanisms. This hybrid approach leverages both the structure- and sequence-aware modeling capabilities, enabling the network to learn more expressive and robust feature embeddings. The proposed model outperformed all baselines in terms of precision, recall, F1 score, and accuracy, confirming its superior ability to model complex behavioral patterns.

4.3. Evaluation.

4.3.1. Metrics. The performance of all models was evaluated using four metrics.

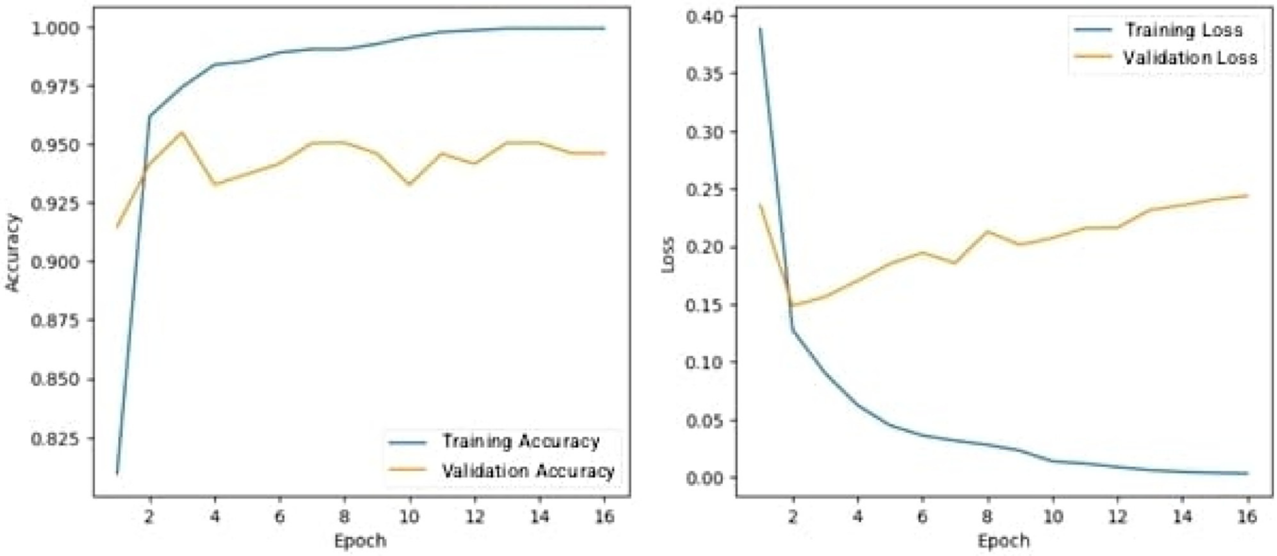

Figure 7. Accuracy and loss curve of GNN (seed 123)

1. **Precision**: The ratio of true positives to predicted positives.

$$Precision = \frac{TP}{TP + FP}$$

Here, TP = true positives and FP = false positives.

2. **Recall**: The ratio of true positives to actual positives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Here, FN = false negatives.

3. **Macro-F1 Score**: Harmonic mean of precision and recall across both classes.

$$Macro\ F1\ Score = 2 \times \frac{Precision + Recall}{Precision \times Recall}$$

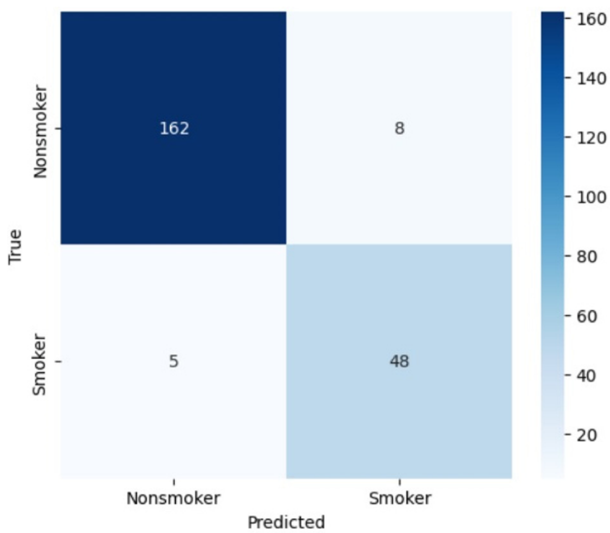

Figure 8. Confusion matrix of FT-transformer (seed 0)

For macro averaging, the F1 Score was computed independently for each class and then averaged.

$$Macro\ F1 = \frac{1}{N}\sum_{i=1}^{N} F1_i$$

Here, N = number of classes and $F1_i = Macro\ F1$ score for class i

4. **Accuracy**: The overall proportion of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

4.3.2. Model Comparison (Train-Test Split). Table 4 presents the overall evaluation results of the test sets of the explored models.

As shown, the proposed GNN + FT transformer outperformed all other models across all evaluation metrics. The Bi-GRU and Bi-LSTM models also achieved high scores but still outperformed in both precision and recall. Traditional ML models, such as *random forest* and *naïve Bayes*, performed moderately well but could not fully capture complex feature interactions.

4.3.3. K-Fold Cross Validation and Ablation Study. To assess the generalization capability of the proposed model more reliably, we applied five-fold cross-validation. This technique helps to mitigate overfitting and provides a better estimate of the model performance on unseen data by ensuring that each data point is in the validation set. To enhance the reliability of our results and account for randomness in model initialization and data splitting further, we repeated the entire five-fold cross-validation procedure using three random seeds: 0, 42, and 123. This setup ensures that our performance metrics are not dependent on a single data split and allows for a more robust and statistically sound evaluation. For each seed, we computed the mean performance across five folds, and reported the overall mean and 95% confidence intervals across all three seeds.

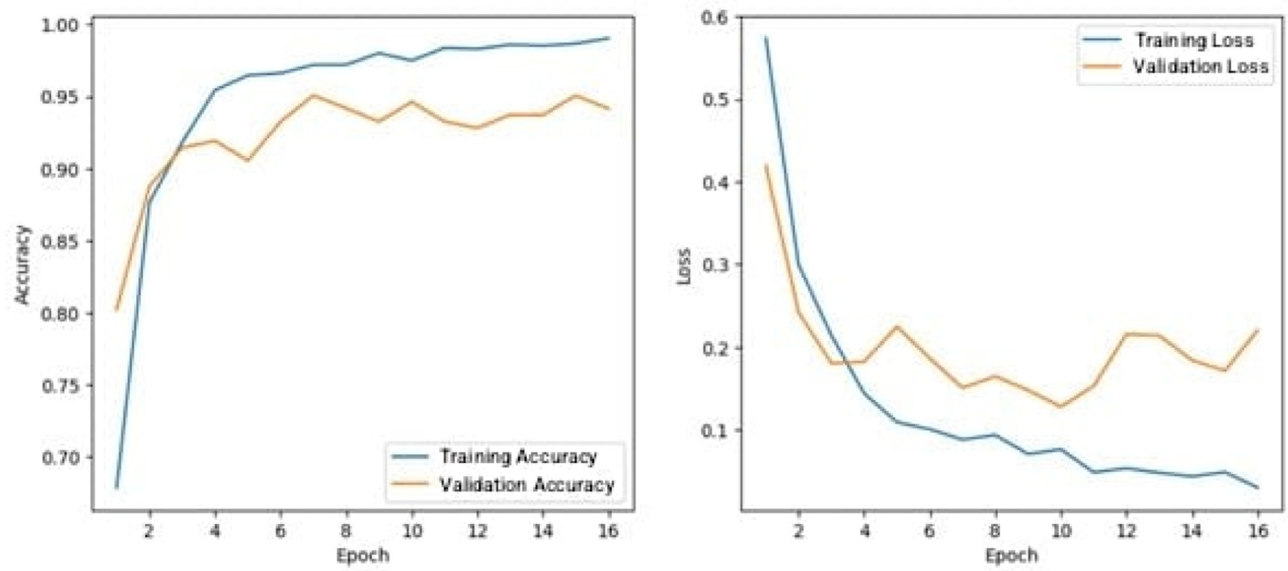

Figure 9. Accuracy and loss curve of the FT-transformer (seed 0)

In addition, an ablation study was conducted to understand the contributions of each component of the proposed architecture. We evaluated and compared the performances of three configurations: GNN, FT-transformer, and GNN + FT-transformer. Table 5 summarizes the results of cross-validation and ablation studies for the three seeds.

5. DISCUSSION

5.1. Train-Test Split. The proposed GNN + FT-Transformer model achieved the highest performance across all key metrics, including 93% accuracy, 0.89 precision, 0.96 recall, and a macro-F1 score of 0.92. Traditional ML models, such as *random forest* and *naïve Bayes*, performed reasonably well but had lower recall, indicating reduced sensitivity to the smoker class.

DL models such as Bi-GRU and Bi-LSTM showed improvements: Bi-GRU achieved up to 89% accuracy; however, they were still outperformed by the proposed hybrid model. As shown in Figure 3, the confusion matrix confirms a balanced classification of the model with fewer false positives and negatives. The receiver operating characteristics (ROC) curve in Figure 4, which has an area under the curve (AUC) of 0.97, further highlights the strong discriminative capability.

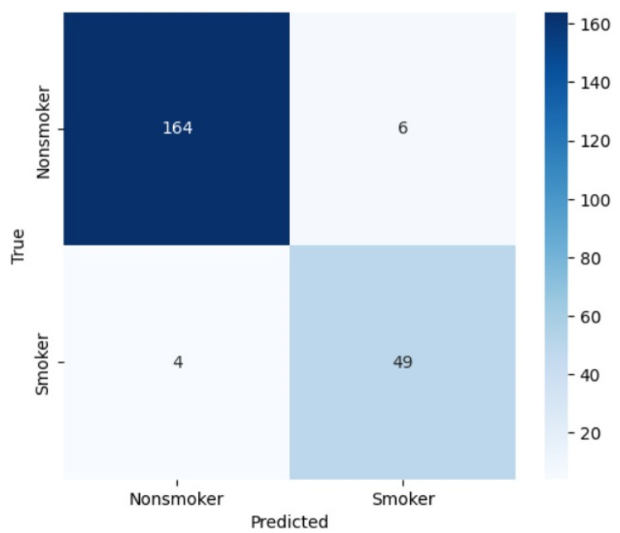

Figure 10. Confusion matrix of GNN + FT-transformer (seed 0)

Qualitatively, the model benefits from the complementary strengths of the GNN and FT-transformer. The GNN component captures the structural relationships among features, such as family and peer influence or mental health, by modeling them as a fully connected graph.

The FT-transformer adds contextual depth by focusing on the most informative features in each instance. This combination helps to manage class imbalances and irregular feature distributions. As shown in Figure 5, the training-validation loss curve shows smooth convergence with no major overfitting. Additionally, the attention weights from both the GNN and transformer modules offer interpretability, consistent with known behavioral patterns of smoking.

5.2. Cross Validation (Multiple Random Seed).

Graph Neural Network (GNN). To evaluate the standalone effectiveness of the GNN, we trained the model using five-fold cross-validation with three different random seeds: 0, 42, and 123. Of these configurations, the model initialized with seed 123 achieved the highest performance.

The confusion matrix in Figure 6 for seed 123 shows how well the GNN could distinguish between classes with minimal misclassifications. Figure 7 shows the training-validation accuracy and loss curve for seed 123. This standalone GNN model served as the baseline for ablation analysis. Its performance validates the ability of the GNN to capture topological relationships and structural dependencies within the data effectively.

FT Transformer. To isolate the contribution of feature-based attention mechanisms, we evaluated the FT-transformer independently without incorporating any graph-structural information. We conducted five-fold cross-validation with three different random seeds: 0, 42, and 123. Of these, the model trained

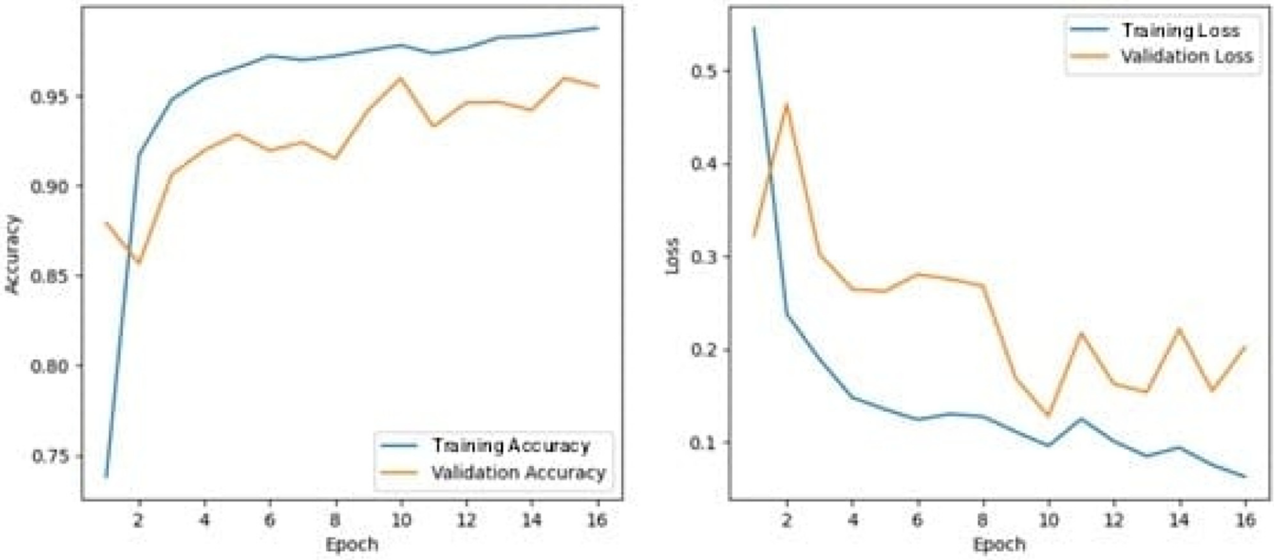

Figure 11. Accuracy and loss curve of GNN +FT-transformer (seed 0)

with seed 0 consistently achieved the highest average accuracy during cross-validation. The confusion matrix shown in Figure 8 further demonstrates the ability of the FT-transformer to model complex patterns in tabular data effectively using its feature-wise attention mechanism. Figure 9 shows the training-validation accuracy and loss curves for seed 0, reflecting the stable convergence of the model. The FT-transformer effectively captures complex feature interactions in the tabular data using its attention-based mechanism, offering strong standalone performance. However, it cannot model inter-entity relationships, limiting its effectiveness in tasks for which structural dependencies are critical.

GNN + FT-Transformer. In the final stage of our ablation study, we evaluated the hybrid GNN + FT-transformer model, which integrated the strengths of both GNNs and FT-transformers to capture relational and feature-level patterns simultaneously. We conducted five-fold cross-validation with three random seeds: 0, 42, and 123. The hybrid model achieved the best performance with seed 0, attaining the highest mean accuracy and showing strong generalization across the validation folds. The confusion matrix in Figure 10 for seed 0 shows highly accurate predictions with minimal misclassifications, indicating that the hybrid model effectively learned from both the graph structure and tabular features. The training–validation curves illustrated in Figure 11 show smooth convergence with no significant overfitting, further validating the stability and robustness of the hybrid approach.

6. CONCLUSIONS

Herein, we have proposed a hybrid GNN + FT-transformer model to classify smoking behavior using sociodemographic, psychological, and behavioral features. The model outperformed traditional ML and DL baselines, achieving an accuracy of 93% and 0.92 macro-F1 score. The success of the model stems from the ability of the GNN to capture feature relationships and the strength of the transformer in the contextual representation. To ensure robustness and reliability, the model was evaluated using a stratified five-fold cross-validation, which consistently demonstrated stable performance across cross-validation runs, thereby mitigating the risk of overfitting and variance arising from data splits. Further evaluation using a confusion matrix, ROC curve, and loss analysis confirmed the robustness and generalization of the model. The approach also offers interpretability and was consistent with known behavioral factors. This study highlights the advantages of hybrid architectures for behavioral prediction and public health applications. A key limitation of the current hybrid model is its dependence on predefined graph structures, which may not accurately capture complex or evolving relationships in the data. Future studies will focus on enabling automatic graph construction and exploring adaptive architectures to improve generalization, particularly in real-world settings with irregular or domain-shifted data.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Sk. Md Abir Hasan Imran – *Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Saidpur 5311, Bangladesh;* 0009-0008-3572-3897
Email: researchabir15@gmail.com

Author

Arupa Barua – *Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Bangladesh;* 0009-0004-5261-225X
Email: arupab362@gmail.com

Corresponding Author

Md. Osama – *Research Mentor, Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology, Saidpur 5311, Bangladesh;* 0009-0001-5307-9993
Email: osama_cse@baust.edu.bd

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to the Department of Computer Science and Engineering, Bangladesh Army University of Science and Technology (BAUST), Saidpur Cantonment, Nilphamari, Bangladesh, for providing continuous support, valuable guidance, and necessary resources throughout this research.

REFERENCES

(1) Ali, N., Paul, Y., Husain, A. & Hussain, H. in *The International Conference on Recent Innovations in Computing.* 93–116 (Springer).

(2) Carney, R. *et al.* The clinical and behavioral cardiometabolic risk of children and young people on mental health inpatient units: A systematic review and meta-analysis. *General Hospital Psychiatry* **70**, 80–97 (2021).

(3) Khan, A., Elhassan, M. A., Khan, S. & Deng, H. Deep learning-based smoker classification and detection: An overview and evaluation. *Expert Systems with Applications*, 126208 (2024).

(4) Ortis, A., Caponnetto, P., Polosa, R., Urso, S. & Battiato, S. A report on smoking detection and quitting technologies. *International journal of environmental research and public health* **17**, 2614 (2020).

(5) Khan, A. A., Laghari, A. A. & Awan, S. A. Machine learning in computer vision: A review. *EAI Endorsed Transactions on Scalable Information Systems* **8** (2021).

(6) Melchior, M., Chastang, J.-F., Mackinnon, D., Galéra, C. & Fombonne, E. The intergenerational transmission of tobacco smoking—the role of parents' long-term smoking trajectories. *Drug and alcohol dependence* **107**, 257–260 (2010).

(7) Becker, T. D., Arnold, M. K., Ro, V., Martin, L. & Rice, T. R. Systematic review of electronic cigarette use (vaping) and mental health comorbidity among adolescents and young adults. *Nicotine and Tobacco Research* **23**, 415–425 (2021).

(8) Mahabee-Gittens, E. M., Xiao, Y., Gordon, J. S. & Khoury, J. C. The dynamic role of parental influences in preventing adolescent smoking initiation. *Addictive behaviors* **38**, 1905–1911 (2013).

(9) Mays, D. *et al.* Parental smoking exposure and adolescent smoking trajectories. *Pediatrics* **133**, 983–991 (2014).

(10) Scalici, F. & Schulz, P. J. Parents' and peers' normative influence on adolescents' smoking: results from a Swiss-Italian sample of middle schools students. *Substance abuse treatment, prevention, and policy* **12**, 1–9 (2017).

(11) Fluharty, M., Taylor, A. E., Grabski, M. & Munafò, M. R. The association of cigarette smoking with depression and anxiety: a systematic review. *Nicotine & Tobacco Research* **19**, 3–13 (2016).

(12) De Luna, R. G. *et al.* in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS).* 84–89 (IEEE).

(13) Ali, N., Paul, Y., Husain, A. & Hussain, H. in *Proceedings of International Conference on Recent Innovations in Computing.* 93 (Springer Nature).

(14) Jarvis, M., Tunstall-Pedoe, H., Feyerabend, C., Vesey, C. & Saloojee, Y. Biochemical markers of smoke absorption and self reported exposure to passive smoking. *Journal of Epidemiology & Community Health* **38**, 335–339 (1984).