AI-Driven ADHD Prediction and Analysis at Early Age:

A Novel Approach Integrating Machine Learning, Explainable AI, LLMs, and Dialogflow with a Virtual Therapist Chatbot (ComfortChat)

Results and Discussion

Senior Design Project

Md. Adham Wahid ID# 2111177042
Fatema Afsan Ema ID# 2022281642
Istiak Ahasan ID# 2012082642

Faculty Advisor:
Dr. Sifat Momen
Associate Professor
ECE Department

Department of Electrical and Computer Engineering
North South University
Spring 2025

Experiment Result Analysis and Discussion

0.1 Results and Discussion

This section illustrates the results of the experiments and comprehensive analysis of the proposed system. Traditional evaluation metrics such as accuracy, precision, recall, and f1 score have been used to measure the performance and findings of the basic, ensemble and complex models for the proposed system. Additionally, for the large language models, including accuracy, precision, recall, and f1 score other metrics like validation loss, runtime , auc has been considered. The confusion matrix, Receiver Operating Characteristic (ROC) and (AOC) curve, and precision–recall curve (PRC), learning and comparsion curves add transparency. Ablation studies, generalization and robustness and domain expert validation brings novelty and clarity to the proposed study. Moreover, hyperparameter analysis and model interpretation have been conducted to provide further insights.

Performance metrics–Accuracy, Precision, Recall, and F1-Score equations has been used as given below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$F1\text{-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Table 1: Definitions of terms used in metrics equations

Term	Definition
TP	True Positive (Correctly predicted positive outcomes)
TN	True Negative (Correctly predicted negative outcomes)
FP	False Positive (Incorrectly predicted positive outcomes)
FN	False Negative (Incorrectly predicted negative outcomes)

0.1.1 Baseline performance (without resampling)

The baseline without applying any resampling techniques provide unbiased and impartial outcomes. Fourteen ML model algorithms have been trained without applying any resampling methods.

Table 2: Performance Metrics of Machine Learning and Other Models

Type	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Basic ML Models	Logistic Regression	92.54 ± 0.21	82.23 ± 0.69	75.34 ± 0.78	78.22 ± 0.70
	Decision Trees	91.34 ± 0.29	77.41 ± 0.75	77.31 ± 0.91	77.36 ± 0.80
	Linear SVC	92.50 ± 0.23	82.63 ± 0.88	74.21 ± 0.62	77.58 ± 0.67
	k-NN	92.32 ± 0.12	81.31 ± 0.54	75.31 ± 0.47	77.87 ± 0.31
	Naïve Bayes	85.72 ± 0.39	69.31 ± 0.41	82.00 ± 0.37	72.98 ± 0.49
Ensemble Models	Random Forest	93.29 ± 0.15	83.36 ± 0.50	79.99 ± 0.80	81.54 ± 0.50
	AdaBoost	92.46 ± 0.18	81.80 ± 0.62	75.60 ± 0.78	78.24 ± 0.61
	XGBoost	93.80 ± 0.13	85.01 ± 0.70	81.05 ± 0.41	82.86 ± 0.14
	LightGBM	93.79 ± 0.09	86.00 ± 0.61	79.24 ± 0.55	82.14 ± 0.23
	Gradient Boosting	93.23 ± 0.07	84.35 ± 0.42	77.49 ± 0.54	80.40 ± 0.27
	CatBoost	93.59 ± 0.06	85.50 ± 0.51	78.49 ± 0.55	81.48 ± 0.24
Blending & Stacking	Blending (ML Models)	93.44 ± 0.03	83.84 ± 0.11	81.40 ± 0.11	82.56 ± 0.06
	Stacking	**94.27 ± 0.02**	**87.00 ± 0.15**	**82.07 ± 0.08**	**84.30 ± 0.03**
	Blending (DL Models)	93.49 ± 0.03	83.94 ± 0.06	81.64 ± 0.14	82.73 ± 0.10

Among the fourteen ML models the stacking ensemble comes out with the best performance having 94.27% accuracy, 87.00% precision, 82.07% recall and 84.30% f1 score. Table 2 portrays the baseline performances of the fourteen ML algorithms of the proposed framework.

0.1.2 Performance with oversampling(SMOTE)

Synthetic Minority Oversampling Technique is referred to as SMOTE. Instead of replicating pre-existing cases, this widely used oversampling technique generates synthetic examples. For analyzing the oversampling outcomes all the algorithms have been trained applying SMOTE.

Table 3: Performance Metrics of Machine Learning Models with SMOTE Resampling

Type	**Model**	**Acc (%)**	**Pre (%)**	**Rec (%)**	**F1 (%)**
Basic ML Models	Logistic Regression	86.32	71.00	88.00	75.00
	Decision Trees	90.97	76.00	78.00	77.00
	Linear SVC	85.93	71.00	88.00	75.00
	k-NN	88.98	73.00	85.00	77.00
	Naïve Bayes	84.13	68.00	84.00	72.00
Ensemble Models	Random Forest	92.78	**81.00**	81.00	82.00
	AdaBoost	89.35	74.00	85.00	78.00
	XGBoost	**92.92**	**81.00**	84.00	**83.00**
	LightGBM	91.88	78.00	87.00	82.00
	CatBoost	91.39	77.00	**89.00**	82.00
	Gradient Boosting	92.76	80.00	83.00	82.00
Blending & Stacking	Stacking Model	92.86	**81.00**	82.00	81.00
	Blending Model	92.82	**81.00**	83.00	82.00

From Table 3 it can be stated that the highest accuracy of 92.92% is achieved by xgboost classifier. The top precision score is 81.00% for four different models (xgboost, random forest, stacking and blending). Conversely, the CatBoost classifier achieved the highest recall score of 89.00%. The xgboost also holds the top f1 score of 83.00%.

0.1.3 Performance with undersampling

The undersampling technique makes the unbalanced dataset balanced by reducing the number of samples from the majority class. This technique has been applied to observe the results by different ML algorithms.

Table 4: Performance Metrics of Machine Learning and Other Models with Undersampling

Type	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Basic ML Models	Logistic Regression	85.70	71.00	88.00	75.00
	Decision Trees	88.39	73.00	87.00	77.00
	Linear SVC	85.19	70.00	88.00	74.00
	k-NN	86.20	71.00	87.00	75.00
	Naïve Bayes	84.42	69.00	84.00	72.00
Ensemble Models	Random Forest	87.91	73.00	90.00	78.00
	AdaBoost	86.23	71.00	89.00	75.00
	XGBoost	88.19	73.00	91.00	78.00
	LightGBM	88.24	74.00	91.00	78.00
	CatBoost	87.40	73.00	**92.00**	77.00
	Gradient Boosting	88.19	73.00	91.00	78.00
Blending & Stacking	Stacking	88.88	74.00	91.00	79.00
	Blending	**89.79**	**75.00**	89.00	**80.00**

Table 4 provides the performances of all ML algorithms for the proposed framework. Interestingly, it has been observed that the blending ensemble performs the best with 89.79% accuracy, 75.00% precision, and 80.00% f1 score. However, the highest recall score has been achieved by the catboost classifier of 92.00%.

0.1.4 Hyperparameter Optimization and Analysis

Table 5: Hyperparameter Tuning Results: With Best Parameters

Method	Model	Accuracy (%)	Best Parameters
GridSearch	Logistic Regression	92.63	{model__C: 0.01, model__solver: 'lbfgs'}
	Decision Trees	92.90	{model__max_depth: 10, model__min_samples_split: 5}
	Linear SVC	92.50	{model__C: 1.0, model__kernel: 'linear'}
	k-NN	92.40	{model__n_neighbors: 9}
	Naïve Bayes	85.72	{alpha: 0.1, fit_prior: True}
	Random Forest	93.64	{model__max_depth: 20, model__n_estimators: 500}
	AdaBoost	92.56	{model__learning_rate: 0.1, model__n_estimators: 100}
	XGBoost	93.80	{model__learning_rate: 0.2, model__n_estimators: 100}
	LightGBM	93.86	{model__learning_rate: 0.1, model__n_estimators: 100}
	Gradient Boosting	93.81	{model__learning_rate: 0.2, model__n_estimators: 300}
	CatBoost	93.87	{model__depth: 6, model__iterations: 300, model__learning_rate: 0.1}
	Stacking (ML models)	94.20	{rf_n_estimators: 150, xgb_learning_rate: 0.05, meta_C: 0.03}
	Blending (ML models)	93.90	{log_C: 0.01, knn_n_neighbors: 9, rf_n_estimators: 300}
Optuna	Logistic Regression	92.75	{C: 0.0027, solver: 'lbfgs'}
	Decision Trees	93.23	{max_depth: 9, min_samples_split: 7}
	Linear SVC	92.70	{C: 0.03, kernel: 'linear'}
	k-NN	92.48	{n_neighbors: 15}
	Naïve Bayes	85.92	{alpha: 0.13, fit_prior: True}
	Random Forest	93.92	{n_estimators: 432, max_depth: 20}
	AdaBoost	92.68	{n_estimators: 131, learning_rate: 0.0589}
	XGBoost	93.95	{learning_rate: 0.0912, n_estimators: 261}
	LightGBM	94.06	{learning_rate: 0.0516, n_estimators: 315}
	Gradient Boosting	94.00	{learning_rate: 0.1093, n_estimators: 420}
	CatBoost	94.14	{learning_rate: 0.0374, depth: 7, iterations: 391}
	Stacking (ML models)	94.30	{rf_n_estimators: 200, lgbm_n_estimators: 250, xgb_n_estimators: 500, xgb_learning_rate: 0.0482, xgb_max_depth: 5, meta_C: 0.0357}
	Blending (ML models)	94.00	{log_C: 0.02, knn_n_neighbors: 11, rf_n_estimators: 350}

Here, Table 5 outlines the best outcome of the ML models with the best parameters. Considering the accuracy, clearly the stacking baseline ensemble outshines all other ML

algorithms with quite a margin. Both optuna and gridsearch confirm the statement. Optuna outperforms gridsearch by using intelligent, adaptive sampling strategies that efficiently explore the hyperparameter space, significantly reducing computation time while achieving better model performance. For this reason, the parameters the optuna provided were chosen and utilized for the suggested system. The params are {'xgb_n_estimators': 500, 'xgb_learning rate': 0.0482, 'xgb_max depth': 5, 'meta_C': 0.0357} .

0.1.5 Confusion matrix

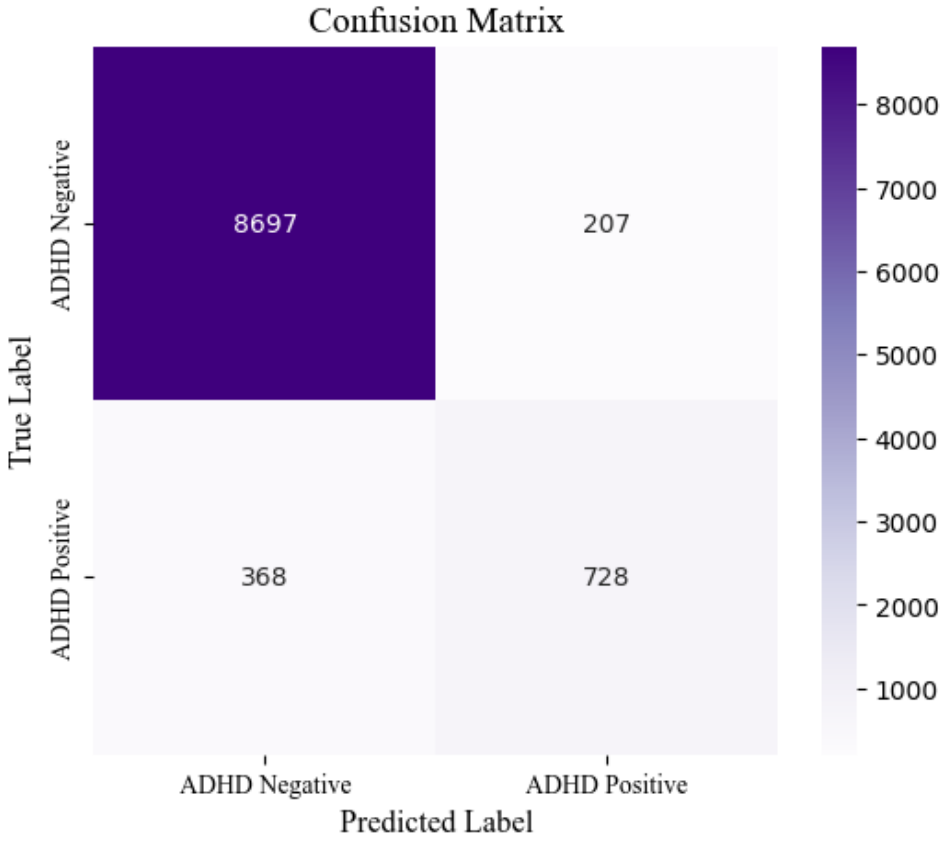

Figure 1: Confusion Matrix for the Proposed Model(Stacking Baseline)

Figure 1 states the confusion matrix for the proposed stacking model. The model holds a strong macro recall of 82.07%, demonstrating the model's generalizability to both classes. The recall for the ADHD-positive cases demonstrates that the model can identify a very high percentage of actual ADHD cases without even employing resampling techniques. Also, the strong recall of the ADHD-negative cases demonstrates that the model is effective at identifying individuals without ADHD, and there are few false positives. As the data is in its raw form, it provides unbiased outcomes.

0.1.6 ROC AUC curves

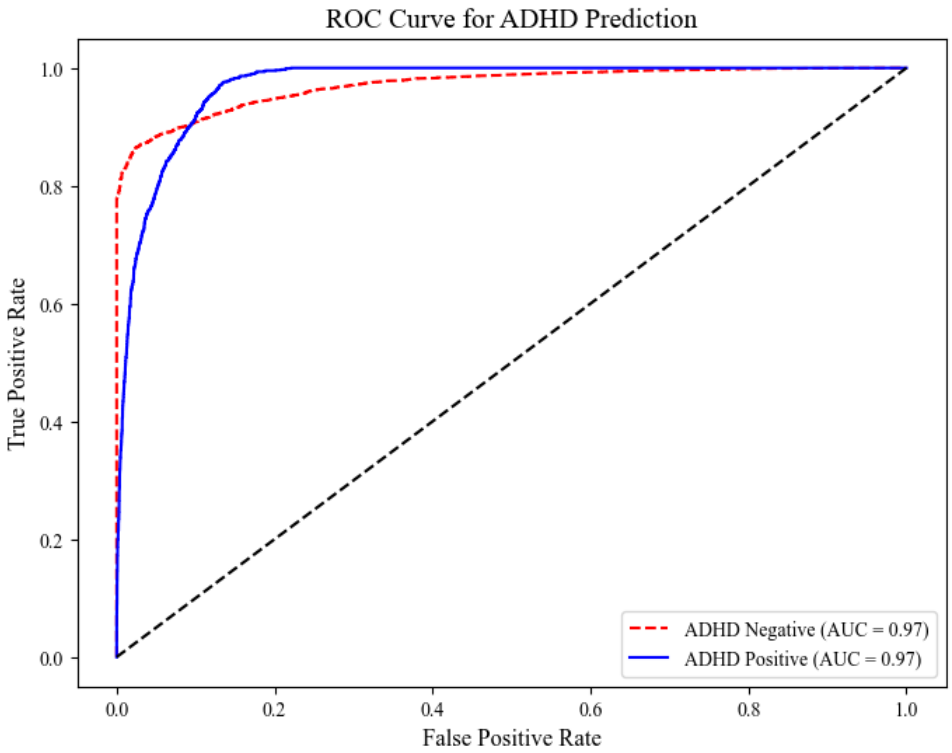

Figure 2: ROC-AUC curve for the Proposed Model(Stacking Baseline)

The ROC-AUC curve from Figure 2 evaluates the performance of the stacking baseline model in distinguishing between ADHD-positive and ADHD-negative cases. Both classes achieve a score of 0.97 which indicates excellent discriminatory ability of the model. The model's durability is demonstrated by its strong performance in both classes, even in the absence of resampling. The ROC curve for both classes is sharply increasing at the beginning, which means that the model has high true positive rates with low false positives, which is desirable for classification problems. AUC values close to 1.0 mean that the model is good at distinguishing between positive and negative cases.

0.1.7 Precision Recall curves

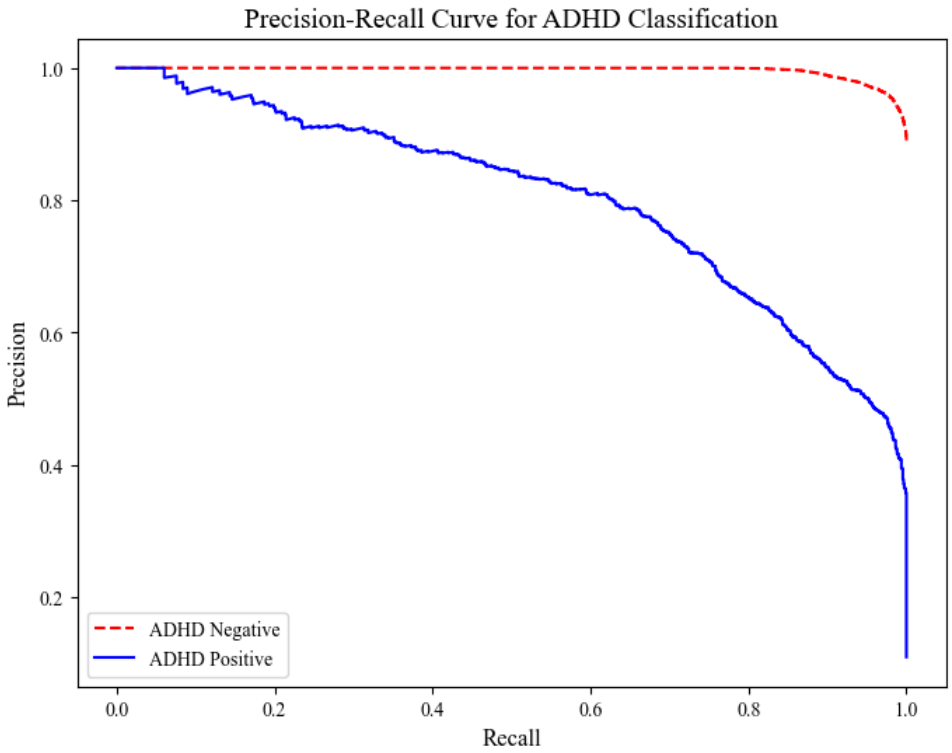

Figure 3: Precision-Recall curve for the Proposed Model(Stacking Baseline)

The Precision-Recall (PR) curve provides a detailed view of how well the stacking ensemble balances precision and recall for ADHD classification. Figure 3 outlines the precision-recall curve for the proposed framework. It has been observed that the model gains 87.00% precision and 82.07% recall values. Both positive and negative class outcomes are well-balanced and maintain a strong overall performance without resampling techniques.

0.1.8 Learning curves

The learning curve provides valuable insights into the stacking baseline model's performance as the training size increases, helping to assess bias-variance tradeoffs and model generalization.

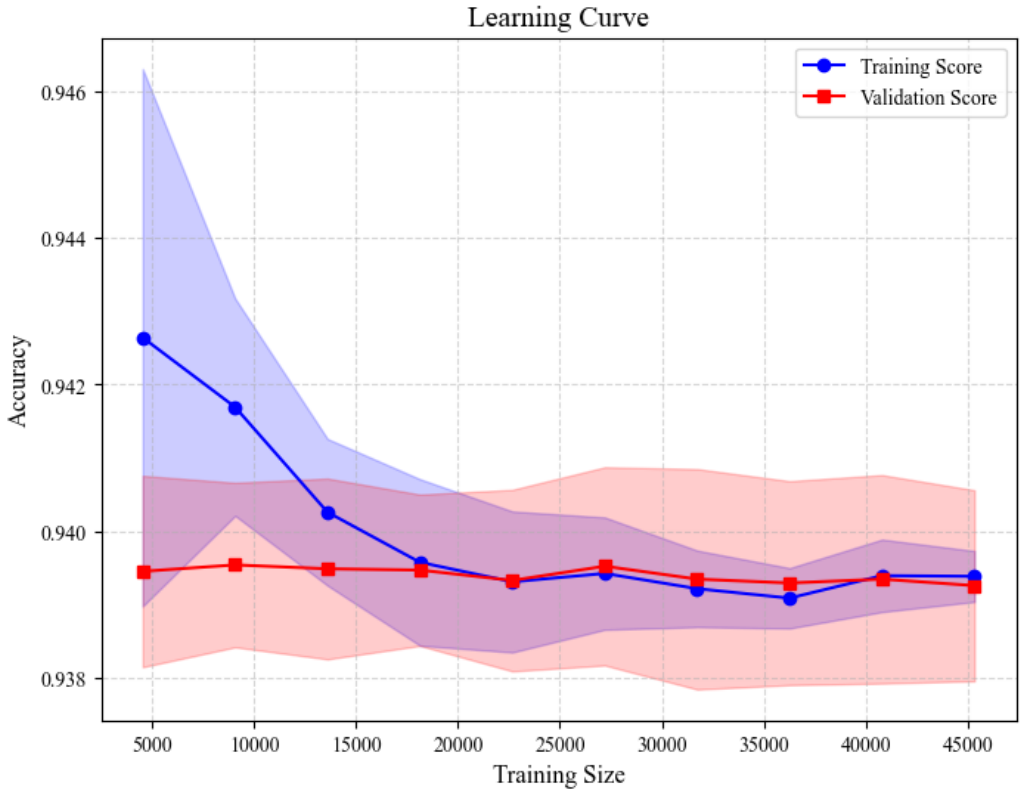

Figure 4: Learning curve for the Proposed Model(Stacking Baseline)

Initially, the model shows high variance as it is working on the smaller datasets but as more and more data flow into the system, the variance reduces while the validation accuracy levels off 94.00%. Not having a huge difference between validation and training scores indicates that neither the model is overfitting nor underfitting, and it is best optimized for the classification of ADHD and the Figure 4 outlines that.

0.1.9 Ablation studies

Ablation experiments [1] remove or modify systematically some features or elements of a model to quantify their contribution separately to performance. Ablation is employed to identify the most impactful features and enhance model complexity in the optimal manner.

Table 6 shows the ablation study results for the proposed study. Removing 5 less important feature based on feature importance provides the best outcome. It can be observed that the performance gradually declines as the number of features decreases.

Table 6: Performance metrics of Machine Learning and other models with feature ablation

Num. of Features	Removed Features	Accuracy	Precision	Recall	F1 Score
20	5	**0.9423**	**0.8674**	**0.8144**	**0.8381**
25	0	0.9417	0.8652	0.8135	0.8367
24	1	0.9413	0.8647	0.8114	0.8352
23	2	0.9419	0.8665	0.8130	0.8369
22	3	0.9416	0.8656	0.8122	0.8361
21	4	0.9423	0.8675	0.8141	0.8380
19	6	0.9407	0.8638	0.8083	0.8330
18	7	0.9390	0.8598	0.8022	0.8277
17	8	0.9390	0.8606	0.8010	0.8272
16	9	0.9370	0.8545	0.7953	0.8213
15	10	0.9365	0.8517	0.7959	0.8206
14	11	0.9361	0.8498	0.7963	0.8201
13	12	0.9360	0.8507	0.7941	0.8190
12	13	0.9344	0.8452	0.7904	0.8146
11	14	0.9322	0.8431	0.7760	0.8047
10	15	0.9294	0.8385	0.7603	0.7926

0.2 Performance evaluation metrics(LLMs)

0.2.1 Transformer Model Evaluation Measures(BERT)

BERT transformer-based models, designed for NLP and highly prevailing NLU, form the foundation architecture for many powerful Large Language Models (LLMs) by enabling deep contextual human language comprehension. In this particular study, five different BERT-based transformed models have been trained for comprehensive experiments. It consists of BERT, DistilBERT, ClinicalBERT, BioBERT and ALBERT models. These models were selected for their respective strengths in general-purpose understanding, lightweight efficiency, and domain-specific expertise in clinical and biomedical texts. Each model was fine-tuned on the task-specific dataset to evaluate performance across key metrics. The comparative results from these variations help assess which adaptation of BERT is most effective for the targeted healthcare-related predictions. These findings provide insights into the trade-offs between model size, domain relevance, and performance. Such analysis also guides the selection of optimal models for real-world deployment in resource-constrained or specialized environments.

Table 7: Evaluation Metrics of Transformer Models (BERT)

Model	Validation Loss	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	AUC	Runtime (s)
BERT	0.247	87.15	91.52	87.15	88.62	0.911	36.47
Epoch 1	0.247	87.14	91.52	87.14	88.62	0.910	-
Epoch 2	0.235	89.19	79.55	89.19	84.10	0.817	-
Epoch 3	0.237	89.19	79.55	89.19	84.10	0.919	-
DistilBERT	0.142	93.45	93.22	93.45	93.32	0.968	19.25
Epoch 1	0.158	92.78	93.20	92.78	92.96	0.960	-
Epoch 2	0.162	93.34	92.78	93.34	92.77	0.965	-
Epoch 3	0.142	93.45	93.22	93.45	93.32	0.968	-
ClinicalBERT	0.343	89.27	90.42	89.27	84.28	0.291	36.04
Epoch 1	0.343	89.27	90.42	89.27	84.28	0.291	-
Epoch 2	0.343	89.19	79.55	89.19	84.10	0.893	-
Epoch 3	0.341	89.19	79.55	89.19	84.10	0.913	-
BioBERT	0.344	89.19	79.55	89.19	84.10	0.731	42.17
Epoch 1	0.344	89.19	79.55	89.19	84.10	0.731	-
Epoch 2	0.252	89.19	79.55	89.19	84.10	0.902	-
Epoch 3	0.236	89.19	79.55	89.19	84.10	0.909	-
ALBERT	0.183	92.70	92.21	92.70	92.38	0.950	46.25
Epoch 1	0.240	87.14	91.52	87.14	88.61	0.905	-
Epoch 2	0.224	92.39	91.75	92.39	91.94	0.911	-
Epoch 3	0.182	92.70	92.21	92.70	92.38	0.950	-

Table 7 outlines the evaluation metrics of the BERT-based transformer models. Each transformer model has been evaluated over the course of three training epochs, allowing for a comprehensive assessment of performance progression across training iterations.

Table 8: Final Comparison of Transformer Models

Model	Loss	Acc (%)	Pre (%)	Rec (%)	F1 (%)	AUC	Runtime (s)
BERT	0.247	87.15	91.52	87.15	88.62	0.911	36.47
DistilBERT	**0.142**	**93.45**	**93.22**	**93.45**	**93.32**	**0.968**	**19.25**
ClinicalBERT	0.343	89.27	90.42	89.27	84.28	0.291	36.04
BioBERT	0.344	89.19	79.55	89.19	84.10	0.731	42.17
ALBERT	0.183	92.70	92.21	92.70	92.38	0.950	46.25

Summarizing the Table 7 , the Table 8 has been retrieved that outlines the final hypertuned optimized results. It also confirms that DistilBERT provides the best outcome with a loss of 0.142, runtime of 19.25s, and accuracy of 93.45%. Also has a high precision, recall and f1 score.

0.2.2 Comparison graphs for BERT models

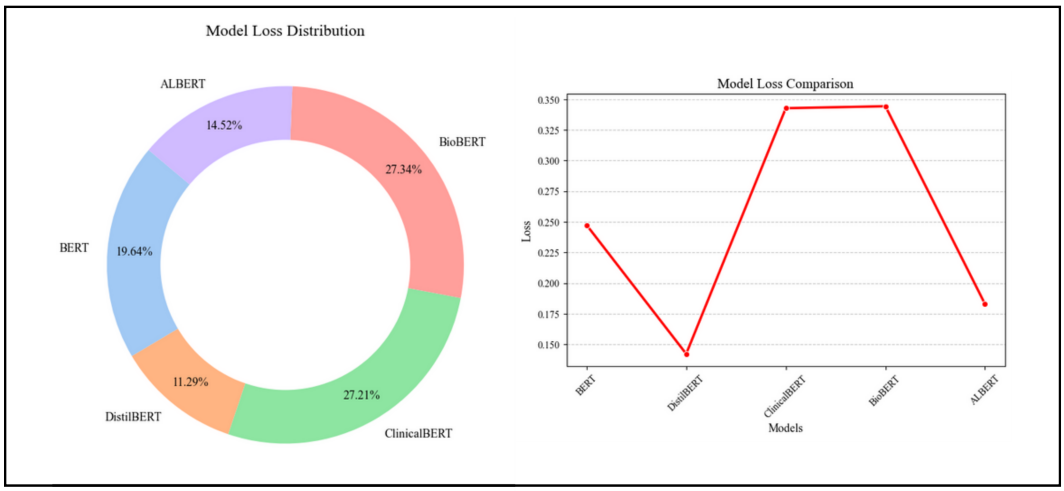

Figure 5: Model Loss Distribution(LLMs)

Analyzing Figure 5 , it can be stated that DistilBERT has the lowest validation loss, according to the model loss distribution, indicating better generalization and optimization. ClinicalBERT and BioBERT, on the other hand, show the greatest losses, indicating comparatively poorer performance across the models that were assessed.

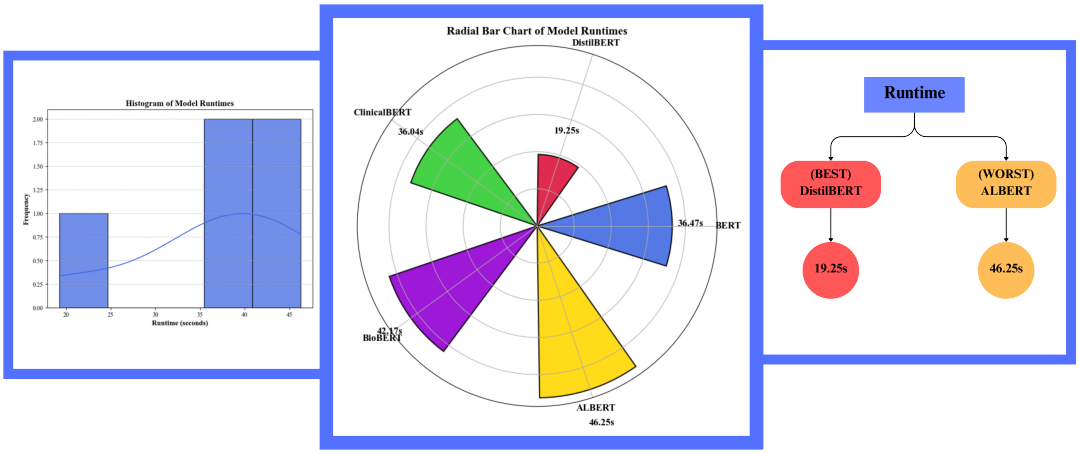

Figure 6: Runtime Distribution(LLMs)

According to the runtime distribution Figure 6, ALBERT takes the longest time (46.25 seconds), and DistilBERT is the most computationally efficient model with the shortest time (19.25 seconds). This shows DistilBERT's computational efficiency advantage without greatly sacrificing performance.

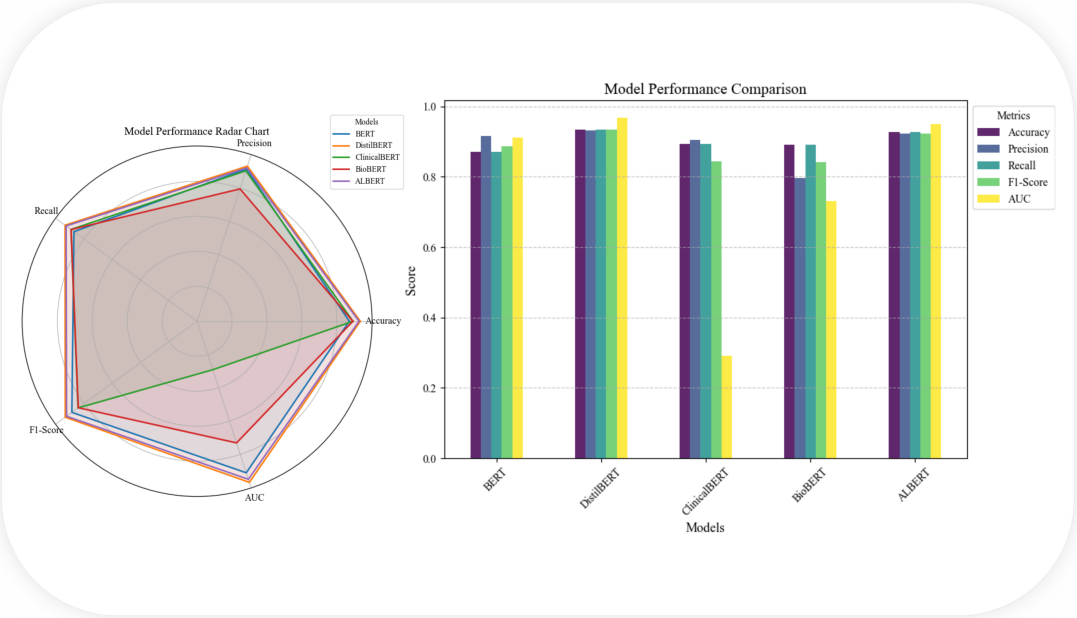

Figure 7: Model Comparison with Radarchart (LLMs)

DistilBERT continues to outperform other models in all the metrics like accuracy, precision, recall, F1-score, and AUC, as can be seen from the radar and bar plots of Figure7. ClinicalBERT's overall performance curve is impacted by its enormous AUC gap. In important metrics, BERT and ALBERT's performance is on par with DistilBERT, but marginally worse.

0.2.3 Generalization and robustness

Performance on 2018–2019 National Survey of Children's Health Dataset

The proposed system was evaluated using the 2018–2019 NSCH dataset [2], with five features removed for alignment. Table 9 presents the results, where the stacking ensemble model achieved the highest accuracy and precision.

Table 9: Performance on 2018–2019 NSCH Dataset (Test Set 1)

Type	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Basic ML Models	Logistic Regression	92.57 ± 0.18	82.33 ± 0.55	75.61 ± 0.57	78.44 ± 0.57
	Decision Trees	91.57 ± 0.11	78.42 ± 0.35	78.01 ± 0.44	78.16 ± 0.48
	Linear SVC	92.51 ± 0.17	82.62 ± 0.62	74.48 ± 0.46	77.77 ± 0.50
	k-NN	92.58 ± 0.18	81.85 ± 0.57	76.86 ± 0.81	79.06 ± 0.63
	Naïve Bayes	86.24 ± 0.38	69.90 ± 0.49	**83.31 ± 0.67**	73.66 ± 0.59
Ensemble Models	Random Forest	93.36 ± 0.18	83.51 ± 0.57	80.47 ± 0.59	81.89 ± 0.50
	AdaBoost	92.54 ± 0.16	81.99 ± 0.50	76.09 ± 0.55	78.63 ± 0.49
	XGBoost	93.73 ± 0.13	84.60 ± 0.47	81.44 ± 0.43	82.92 ± 0.34
	LightGBM	93.91 ± 0.13	85.21 ± 0.51	81.69 ± 0.37	**83.32 ± 0.31**
	Gradient Boosting	93.68 ± 0.12	85.01 ± 0.53	80.21 ± 0.73	82.36 ± 0.43
	CatBoost	93.67 ± 0.18	85.47 ± 0.64	79.26 ± 0.70	81.96 ± 0.55
Blending & Stacking	Blending (ML models)	93.44 ± 0.08	84.29 ± 0.16	76.13 ± 0.32	79.49 ± 0.20
	Stacking (ML models)	**94.27 ± 0.08**	**86.08 ± 0.40**	80.16 ± 0.72	82.78 ± 0.38

The stacking model achieved the best accuracy (94.27%) and precision (86.08%), showcasing excellent performance and robustness during generalization on past-year survey data.

Performance on 2022–2023 National Survey of Children's Health Dataset

Further evaluation was conducted using the latest 2022–2023 NSCH dataset [2]. As shown in Table 10, the blending ensemble model outperformed all others.

Table 10: Performance on 2022–2023 NSCH Dataset (Test Set 2)

Type	Model	Acc (%)	Pre (%)	Rec (%)	F1 (%)
Basic ML Models	Logistic Regression	92.62 ± 0.16	82.50 ± 0.51	75.73 ± 0.44	78.58 ± 0.47
	Decision Trees	91.43 ± 0.21	77.85 ± 0.45	77.04 ± 0.48	77.48 ± 0.44
	Linear SVC	92.51 ± 0.16	82.68 ± 0.63	74.40 ± 0.39	77.73 ± 0.45
	k-NN	92.44 ± 0.19	81.39 ± 0.52	76.59 ± 0.81	78.71 ± 0.67
	Naïve Bayes	86.45 ± 0.39	70.08 ± 0.52	83.22 ± 0.66	73.84 ± 0.61
Ensemble Models	Random Forest	93.04 ± 0.17	82.62 ± 0.61	79.64 ± 0.39	81.03 ± 0.39
	AdaBoost	92.55 ± 0.14	82.06 ± 0.45	76.07 ± 0.56	78.64 ± 0.45
	XGBoost	93.53 ± 0.11	84.17 ± 0.43	80.56 ± 0.57	82.23 ± 0.34
	LightGBM	93.78 ± 0.08	84.98 ± 0.33	81.09 ± 0.54	82.87 ± 0.29
	Gradient Boosting	93.64 ± 0.08	84.81 ± 0.24	80.20 ± 0.84	82.27 ± 0.45
	CatBoost	93.60 ± 0.14	85.18 ± 0.62	79.22 ± 0.65	81.82 ± 0.44
Blending & Stacking	**Blending (ML models)**	**94.98 ± 0.07**	**89.14 ± 0.12**	**86.25 ± 0.33**	**87.62 ± 0.19**
	Stacking (ML models)	93.88 ± 0.18	88.20 ± 0.53	80.73 ± 0.89	83.91 ± 0.63

The blending model surpassed all others, reaching an accuracy of 94.98%, precision of 89.14%, and F1-score of 87.62%. This marks a significant leap in performance over the stacking model, reflecting the proposed system's ability to generalize across independent, unseen datasets.

0.2.4 (Clinical/Domain)Expert assessment and feedback

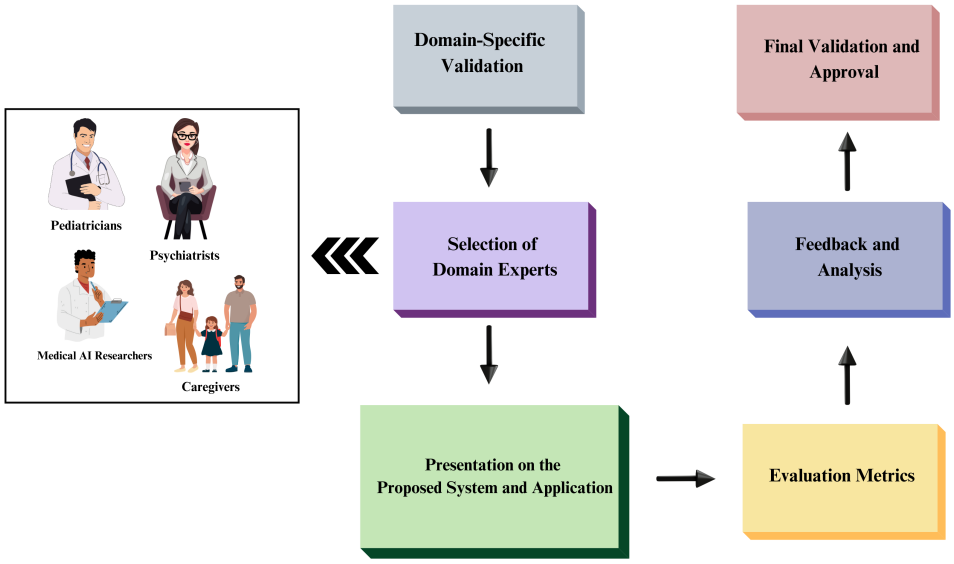

Figure 8: Domain Expert Validation

To add clarity and translucency, we approached various institutions and professionals for raw feedbacks and validations. Initially, the feedbacks and assessments are taken from National Institute of Mental Health (NIMH) & Hospital, Bangladesh[3]. Psychiatrist specializing in child and adolescent mental health, along with registry staff, offered substantive feedback that affirmed the study's promising implications and favorable reception. Another feedback has been taken from one child specialist medical caregiver who is currently working in Enam Medical College, Savar[4]. Moreover, to enhance generalizability further, we collected feedback from a more diverse group of stakeholders including parents, medical caregivers, psychiatrists, and researchers in AI. These comprehensive views also contributed significantly towards validating and refining the system which lead to improvise the system.

0.3 Model Interpretability

Overview of XAI

Explainable AI stands out with the interpretations and insights. They are often responsible for extracting the feature importance. LIME and SHAP have been used for the proposed modelling. These methods help in visualizing how each feature contributes to the final prediction, thereby increasing transparency and trust in the model. Such interpretability is especially crucial in healthcare-related applications where decisions must be justifiable. The insights obtained also support domain experts in understanding key behavioral and medical indicators.

LIME outcomes

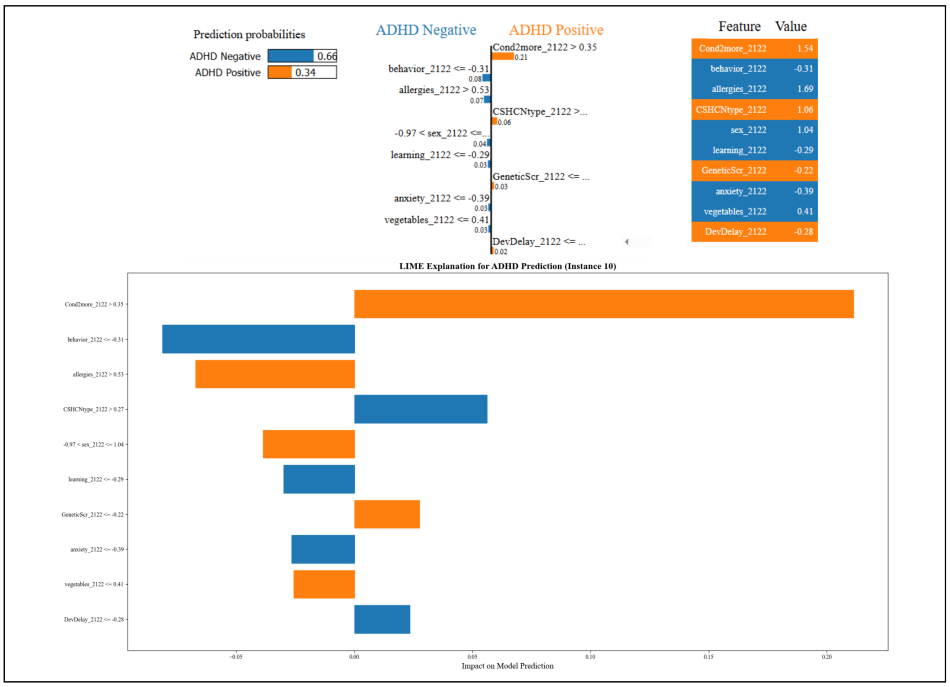

Figure 9: LIME explanation

LIME is a method that uses a local interpretable model to explain every single prediction, which can model any black box machine learning model. The model is estimating a 66% probability of the instance being ADHD Negative and a 34% probability of being ADHD Positive as shown in the Figure 9. The model predicts "ADHD Positive" if the person has more than one health condition (Cond2more_2122), allergies, or a particular type of CSHCN, showing that these are significant risk factors. Behavioral (behavior_2122),

anxiety, eating (vegetables_2122), and development signs (DevDelay_2122, learning_2122) appear to reduce the chances of predicting ADHD in this instance.

Table 11: LIME Explanation for ADHD Prediction (Instance 10)

Feature	**Value**	**Impact on ADHD Prediction**
Cond2more_2122	1.54	Strongest positive influence, significantly increasing probability
allergies_2122	1.69	Moderate positive influence
CSHCNtype_2122	1.06	Mild positive influence
GeneticScr_2122	-0.22	Small positive and slight negative influences
DevDelay_2122	-0.28	Minor positive and slight negative influences
behavior_2122	-0.31	Strongest negative influence, significantly decreasing probability
anxiety_2122	-0.39	Moderate negative influence
vegetables_2122	0.41	Moderate negative influence
learning_2122	-0.29	Slight negative influence

The Table 11 illustrates the LIME Explanation for ADHD Prediction. It also picturize that children having multiple chronic diseases and severe behavioral issues have most impact on ADHD positive and reverse cases for the ADHD negetive.

0.3.1 SHAP outcomes

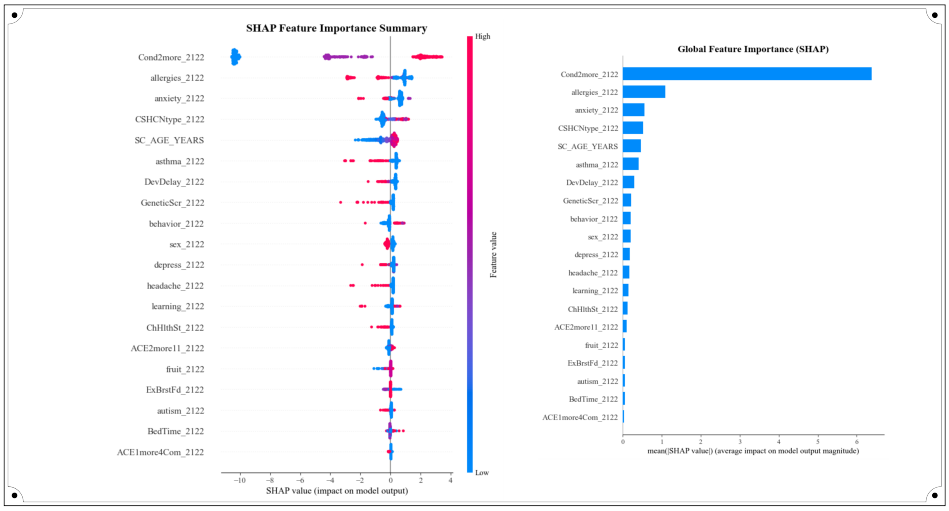

Figure 10: SHAP explanation

Through Figure 10, SHAP also confirms the (Cond2more_2122) is the most influential feature for the ADHD positive cases in children. Its strong and steady correlation with ADHD outcomes is indicated by its high SHAP scores. (allergies_2122,anxiety_2122) also hold substantial global influence. The limited global influence of features such as sex_2122, depress_2122, headache_2122, learning_2122, and others suggests that their influence on ADHD prediction is weaker and less consistent across children.

0.3.2 Decision tree surrogation

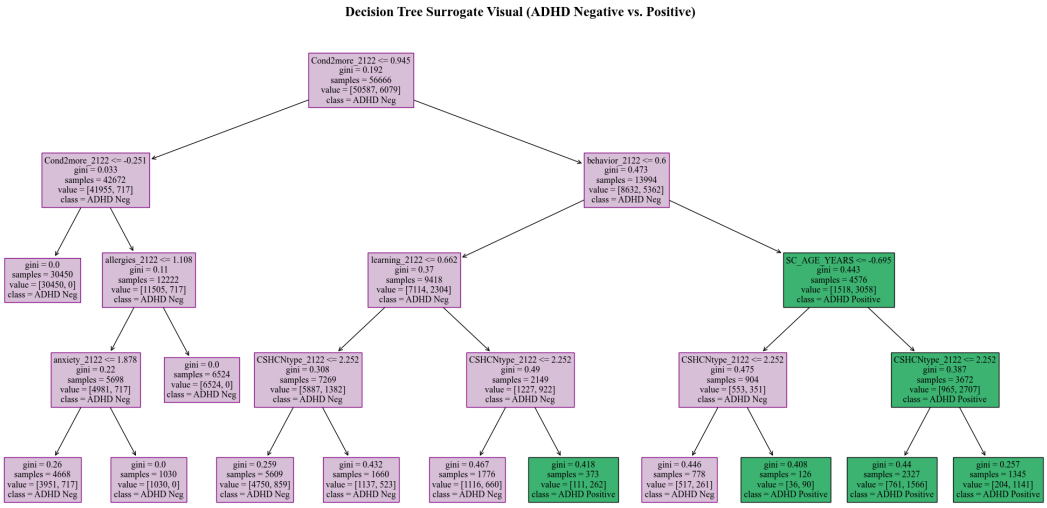

Figure 11: Decision Tree Surrogated Result

To enhance the interpretability of the proposed system for predicting ADHD, a decision tree surrogate has been employed to approximate the decision boundaries learned by the machine learning model like Figure 11. The decision tree provides a structured visualization of feature importance and decision paths and allows for an intuitive comprehension of the reasoning process of the model.

Decision tree proxy analysis confirms (Cond2more_2122) as the optimal predictor of ADHD, with values enhancing ADHD-positive classification. Behavior traits (behavior_2122), anxiety (anxiety_2122), and age (SC_AGE_YEARS) are all significant in the narrowing of predictions. Childs with over one health condition are identified by the model as being at high risk for ADHD. Learning disabilities (CSHCNtype_2122) and special healthcare needs are other case discriminators. These findings are consistent with SHAP and LIME results to enhance the interpretability of ADHD prediction and validate early intervention targeting.

0.4 Deployment and Integration

The proposed AI-Driven ADHD Prediction and Analysis system has been effectively deployed to make the solution accessible and user-friendly in real-world settings. The proposed system has been deployed into both web and mobile application platforms for user convinience.

0.4.1 Web application(Streamlit)

The proposed ensemble stacking model has been deployed into Streamlit[5] for interactive web-based access and user-friendly activities.

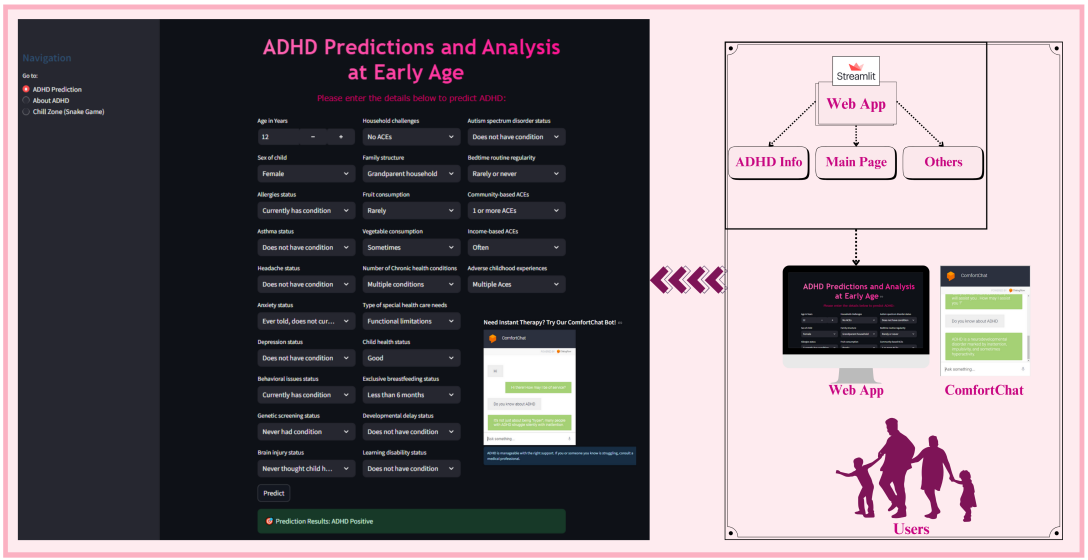

Figure 12: Interface of Webapp

Figure 12 , illustrates the home and prediction page of the developed web application. Moreover, it contains a navigation bar which includes main page, about ADHD page and others page. The chatbot is integrated both in main page and about ADHD page. Users are able to input 25 different questionarries and predict the ADHD outcome.

0.4.2 Mobile application(Gradio)

Apart from deployment in web application , the ensemble stacking model also has been deployed into mobile interface through Gradio[6].

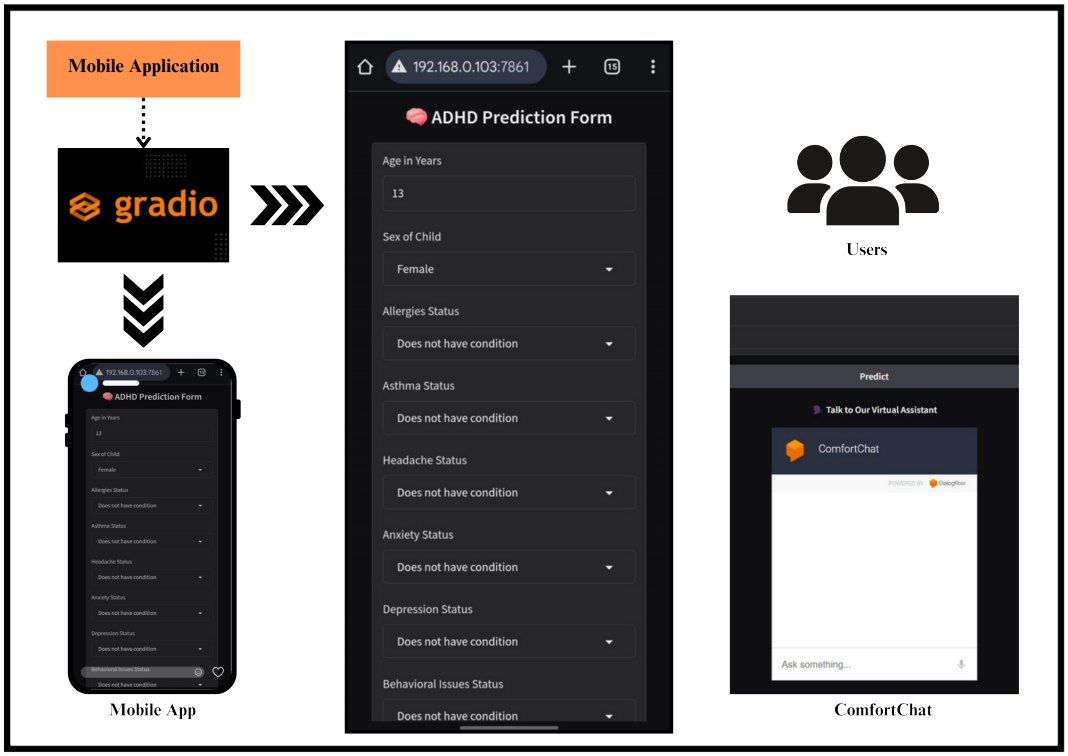

Figure 13: Interface of Mobileapp

The mobile application also kept user-friendly and interactive integrating the chatbot named ComfortChat. Figure 13 portrays the mobile app interface and external looks.

0.5 Discussion

This study introduces a novel and integrative approach for early ADHD detection in children by combining machine learning, large language models (LLMs), explainable AI (XAI), and a therapeutic chatbot. It addresses six core research questions, as outlined in the Introduction section.

For Research Question 1, Table 2 confirms that the proposed stacking ensemble model—built with Random Forest, LightGBM, and XGBoost as base learners and logistic regression as the meta-learner—achieved top performance (accuracy: 94.27%, precision: 87.00%, recall: 82.07%, F1: 84.30%). This clearly validates the strength of ML-based ADHD prediction. Comparison with related works in Table 12 and results from the generalization study (Section 0.2.3) further support the robustness of this approach.

For Research Question 2, interpretability was ensured using LIME, SHAP, permutation importance, and decision tree surrogates (Section 0.3). These XAI methods revealed that chronic illnesses and adverse childhood conditions significantly influence ADHD out-

comes, thereby improving transparency and trust.

For Research Question 3, Table 8 highlights the effectiveness of BERT-based LLMs—especially DistilBERT—which achieved 93.45% accuracy and an AUC of 0.968 while maintaining the lowest runtime. Figure 7 illustrates the efficiency and consistency of transformer models in text-based ADHD analysis.

For Research Question 4, the Dialogflow-based chatbot 'ComfortChat' (Figure : chatbot) was introduced to offer therapeutic aid, accessible through both web and mobile interfaces. It provides users—including caregivers and parents—with preliminary guidance and mental health support in a lightweight and user-friendly format.

Research Question 5 explored how socio-environmental factors, particularly Adverse Childhood Experiences (ACEs) and parental support, impact ADHD. Section: EDA, especially Subsection: ACEs, reveals that ADHD prevalence rises dramatically with ACE exposure—from 8.8% (no ACEs) to 30.03% (with discrimination). Visualizations (Figures: Aces_density and social support_positive) reinforce the role of environmental and household context in ADHD likelihood.

Finally, Research Question 6 validates the system's real-world applicability. Performance on unseen datasets (Tables 9 and 10) shows impressive generalization, with the blending ensemble reaching 94.98% accuracy on the 2022–2023 NSCH dataset—surpassing all prior results. Additionally, expert feedback from institutions like NIMH and Enam Medical College, as well as diverse stakeholders (including parents, clinicians, and AI researchers), confirmed both the system's validity and potential impact in practical deployment scenarios.

To the best of our knowledge, this is the first study to propose a comprehensive solution that combines predictive modeling and a therapeutic chatbot for ADHD support. The integration of generalization on unseen datasets and validation from domain experts enhances both credibility and real-world applicability. These components together form the core novelty of our proposed framework.

Table 12: Comparative analysis of proposed methods on ADHD prediction with similar works

REF	Year	Dataset	Instances	Best Model	Accuracy / AUC
[7]	2023	Custom	342	Decision Tree	82%
[8]	2022	National Survey of Children's Health (NSCH, 2018–2019)	45,779	Random Forest	85.5%
[9]	2021	Custom	50	SVM	76.2%
[10]	2025	ADHD-200 repository	504	Random Forest	96%
[11]	2021	Survey at Kongju National University	5,726	Random Forest	93.6%
[12]	2023	Custom (Japanese school students aged 5–14)	29	Random Forest	93.10%
[13]	2016	Social Responsiveness Scale	2,925	SVM	96.5%
[14]	2025	EEG dataset from Shahed University	100,000	Random Forest	73.71%
[15]	2021	Custom	-	ANN	99.82%
[16]	2023	Custom (Reverse Stroop Task - RST)	243	Random Forest	91.67%
[17]	2023	Custom	973	EEG-based Model	98.2%
[18]	2023	Swedish national registers	238,696	DNN	AUC: 0.684
[19]	2022	Medical records from TU Dresden	299	SVM	66.1%
[20]	2021	Wainstein et al. study (Chile)	50	Voting Ensemble + Meta Learner	AUC: 0.856
[21]	2023	ABCD Study	12,348	LightGBM	AUC: 0.79
[22]	2024	Reddit posts (2011–2017)	Clinical + Non-clinical (Millions)	BERT, RoBERTa, OpenAI GPT	-
[23]	2023	Social networks and organizations	-	BERT	83%
[24]	2025	PRIMATE dataset	1,034 posts	GPT-3.5-Turbo (GAD-7)	95%
This Work	2025	CAHMI 2021–2022 dataset	104,995	Stacking Ensemble, DistilBERT	ML: 94.27%, LLMs: 93.45% , AUC: 0.97

Table 12 shows the comprative analysis of the proposed system with similar studies. Our study achieves state-of-the-art performance in the CAHMI dataset [25] and is optimized using techniques of Machine Learning, LLMs and Dialogflow. The system not only outperforms in terms of predictive accuracy but also offers interpretability through explainable AI components. Its real-time applicability and integration capability with virtual assistants make it practical for deployment in clinical and educational settings. These aspects collectively demonstrate the robustness and adaptability of the framework.

Bibliography

[1] R. Meyes, M. Lu, C. W. de Puiseau, and T. Meisen. Ablation studies in artificial neural networks. *arXiv preprint*, arXiv:1901.08644, 2019.

[2] Johns Hopkins Bloomberg School of Public Health. Cahmi - child and adolescent health measurement initiative. `https://publichealth.jhu.edu/departments/population-family-and-reproductive-health/research-and-practice/life-course-framework/child-and-adolescent-health/cahmi-child-and-adolescent-health-measurement-initiative`, 2025. Accessed: February 2025.

[3] National Institute of Mental Health, Bangladesh. National institute of mental health (nimh) - bangladesh. `https://nimh.gov.bd/`, 2025. Accessed: February 2025.

[4] Enam Medical College and Hospital. Enam medical college and hospital, savar. `https://emch.com.bd/`, 2025. Accessed: February 2025.

[5] Streamlit Inc. Streamlit — the fastest way to build and share data apps. `https://streamlit.io`, 2023. Accessed: February 2025.

[6] Gradio Labs. Gradio: Build machine learning web apps in python. `https://www.gradio.app`, 2023. Accessed: February 2025.

[7] S. Grazioli, A. Crippa, E. Rosi, A. Candelieri, S. B. Ceccarelli, M. Mauri, M. Manzoni, V. Mauri, S. Trabattoni, M. Molteni, P. Colombo, and M. Nobile. Exploring telediagnostic procedures in child neuropsychiatry: addressing adhd diagnosis and autism symptoms through supervised machine learning. *European Child and Adolescent Psychiatry*, 33(1):139–149, January 2024. Epub 2023 Jan 25.

[8] Md. Maniruzzaman, J. Shin, and Md. Al Mehedi Hasan. Predicting children with adhd using behavioral activity: A machine learning analysis. *Applied Sciences*, 12(5):2737, 2022.

[9] W. Das and S. Khanna. A robust machine learning based framework for the automated detection of adhd using pupillometric biomarkers and time series analysis. *Scientific Reports*, 11:16370, 2021.

[10] H. Qin, L. Zhang, J. Wang, et al. Interpretable machine learning approaches for children's adhd detection using clinical assessment data: An online web application deployment. *BMC Psychiatry*, 25:139, 2025.

[11] S. Kim, H. K. Lee, and K. Lee. Can the mmpi predict adult adhd? an approach using machine learning methods. *Diagnostics*, 11(976), 2021.

[12] J. Shin, M. Maniruzzaman, Y. Uchida, M. A. M. Hasan, A. Megumi, and A. Yasumura. Handwriting-based adhd detection for children having asd using machine learning approaches. *IEEE Access*, 11:84974–84984, 2023.

[13] M. Duda, R. Ma, N. Haber, and D. P. Wall. Use of machine learning for the behavioral distinction of autism and adhd. *Translational Psychiatry*, 6(2):e732, 2016.

[14] Nora Fink. A high-accuracy supervised machine learning approach for adhd classification using eeg signals. `https://ssrn.com/abstract=5146511`, 2025. Available at SSRN: `https://ssrn.com/abstract=5146511` or `http://dx.doi.org/10.2139/ssrn.5146511`.

[15] Catherine Joy, Thomas S. George, Albert A. Rajan, and M. S. P. Subathra. Detection of adhd from eeg signals using different entropy measures and ann. *Clinical EEG and Neuroscience*, 53(1):12–23, 2022.

[16] Md. Maniruzzaman, Koki Hirooka, Yoichi Tomioka, Md. Al Mehedi Hasan, Yong Seok Hwang, Akiko Megumi, Akira Yasumura, and Jungpil Shin. Machine learning-based adhd detection from fnirs signal during reverse stroop tasks. *IEEE Access*, 12:82984–82995, 2024.

[17] C. Nash, R. Nair, and S. M. Naqvi. Machine learning in adhd and depression mental health diagnosis: A survey. *IEEE Access*, 11:86297–86317, 2023.

[18] M. Garcia-Argibay, Y. Zhang-James, S. Cortese, et al. Predicting childhood and adolescent attention-deficit/hyperactivity disorder onset: a nationwide deep learning approach. *Molecular Psychiatry*, 28:1232–1239, 2023.

[19] P. Mikolas, A. Vahid, F. Bernardoni, et al. Training a machine learning classifier to identify adhd based on real-world clinical data from medical records. *Scientific Reports*, 12:12934, 2022.

[20] S. Khanna and W. Das. A novel application for the efficient and accessible diagnosis of adhd using machine learning (extended abstract). In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)*, pages 51–54, 2020.

[21] W. P. Kim, H. J. Kim, S. P. Pack, J. H. Lim, C. H. Cho, and H. J. Lee. Machine learning-based prediction of attention-deficit/hyperactivity disorder and sleep problems with wearable data in children. *JAMA Network Open*, 6(3):e233502, 2023.

[22] M. Abdullah and N. Negied. Detection and prediction of future mental disorder from social media data using machine learning, ensemble learning, and large language models. *IEEE Access*, 12:120553–120569, 2024.

[23] Prasenjit Mukherjee et al. Detection of autism spectrum disorder (asd) from natural language text using bert and chatgpt models. *International Journal of Advanced Computer Science and Applications*, 14(10), 2023.

[24] K. Roy, H. Surana, D. Eswaramoorthi, Y. Zi, V. Palit, R. Garimella, and A. Sheth. Exploring the potential of large language models for assisting with mental health diagnostic assessments: The depression and anxiety case, 2025. Preprint.

[25] Johns Hopkins Bloomberg School of Public Health. Child and adolescent health measurement initiative (cahmi). `https://www.jhsph.edu/cahmi`, 2023.